AD_____

Award Number: W81XWH-05-1-0292

TITLE: A Computer-Aided Diagnosis System for Breast Cancer Combining Mammography and Proteomics

PRINCIPAL INVESTIGATOR: Jonathan Jesneck, Ph.D.

CONTRACTING ORGANIZATION: Duke University Medical Center
Durham, NC 27710

REPORT DATE: May 2007

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01/05/07	Annual Summary	1 May 2006 – 30 Apr 2007

4. TITLE AND SUBTITLE

A Computer-Aided Diagnosis System for Breast Cancer Combining Mammography and Proteomics

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-05-1-0292

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Jonathan Jesneck, Ph.D.

E-Mail: jonathan.jesneck@duke.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Duke University Medical Center
Durham, NC 277

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT: This study investigated a computer-aided diagnosis system for breast cancer by combining the following three data sources: mammogram films, radiologist-interpreted BI-RADS descriptors, and proteomic profiles of blood sera. We implemented under 100-fold cross-validation various classification algorithms, including Bayesian probit regression, iterated Bayesian model averaging, linear discriminant analysis, artificial neural networks, as well as a novel method of decision fusion. The top-performing classifier, decision fusion achieved AUC = 0.85 ± 0.01 on the calcification data set and 0.94 ± 0.01 on the mass data set. Decision fusion had a slight performance gain over the ANN and LDA ($p = 0.02$), but was comparable to Bayesian probit regression. Decision fusion significantly outperformed the other classifiers ($p < 0.001$). The blood serum proteins detected lesions moderately well (AUC = 0.82 for normal vs. malignant and normal vs. benign) but failed to distinguish benign from malignant lesions (AUC = 0.55), suggesting they indicate a secondary effect, such as inflammatory response, rather than a role specific for cancer.

15. SUBJECT TERMS
computer-aided diagnosis, digital mammography, clinical proteomics, biopsy, receiver operating characteristic, Bayesian regression, ensemble classifier

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRMC
U	U	U	UU	85	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
15-06-2007	Final	4/2006-4/2007

4. TITLE AND SUBTITLE
A Computer-aided Diagnosis System for Breast Cancer Combining Mammography and Proteomics

5a. CONTRACT NUMBER
W81XWH-05-1-0292

5b. GRANT NUMBER
BC043171

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Jonathan Jesneck, Ph.D.
Joseph Lo, Ph.D.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Duke Advanced Imaging Labs
Duke University Medical Ctr.
2424 Erwin Rd.
Suite 302
Durham, NC 27705

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This study investigated a computer-aided diagnosis system for breast cancer by combining the following three data sources: mammogram films, radiologist-interpreted BI-RADS descriptors, and proteomic profiles of blood sera. We implemented under 100-fold cross-validation various classification algorithms, including Bayesian probit regression, iterated Bayesian model averaging, linear discriminant analysis, artificial neural networks, as well as a novel method of decision fusion. The top-performing classifier, decision fusion achieved AUC = 0.85 ± 0.01 on the calcification data set and 0.94 ± 0.01 on the mass data set. Decision fusion had a slight performance gain over the ANN and LDA ($p = 0.02$), but was comparable to Bayesian probit regression. Decision fusion significantly outperformed the other classifiers ($p < 0.001$). The blood serum proteins detected lesions moderately well (AUC = 0.82 for normal vs. malignant and normal vs. benign) but failed to distinguish benign from malignant lesions (AUC = 0.55), suggesting they indicate a secondary effect, such as inflammatory response, rather than a role specific for cancer.

15. SUBJECT TERMS
computer-aided diagnosis, digital mammography, clinical proteomics, biopsy, receiver operating characteristic, Bayesian regression, ensemble classifier

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
					19b. TELEPHONE NUMBER *(include area code)*

INTRODUCTION

This study investigated a computer-aided diagnosis (CADx) system for breast cancer by combining the following three data sources: mammogram films, radiologist-interpreted BI-RADS descriptors, and proteomic profiles of blood sera.

Although mammography is the modality of choice for early detection of breast cancer[1,2], it has a low positive predictive value (PPV). As a result, only 15 to 34% of women with radiographically-suspicious, nonpalpable lesions are actually found to have a malignancy by histologic diagnosis after biopsy.[3,4] The excessive biopsy of benign lesions raises the cost of mammographic screening[5] and results in emotional and physical burden to the patients, as well as financial burden to society.

In addition to mammography, both BI-RADS descriptors[5] and clinical proteomics[6] have been useful in differentiating benign from malignant breast masses. The combination of mammographic and proteomic information can lead to a more specific classifier for difficult cases. Ensemble classifiers for breast cancer combining multiple sources of information have been shown to outperform classifiers using only one of the information sources.[7]

This research has two purposes. The first is to create three separate classifiers for breast cancer based on proteomic information, mammogram information, and radiologist-interpreted. The second is to combine the outputs of these three first-stage classifiers into one ensemble classifier for breast cancer, which will outperform any of the component classifiers.

Note that although this predoctoral fellowship was awarded for three years, it has now been concluded in two years. The recipient, Jonathan Jesneck, just graduated in May 2007 and has resigned from the fellowship. This fellowship provided him with a solid foundation in cancer research and has allowed him to continue with cancer research at the Dana-Farber Cancer Institute.

BODY

Task 1. Build a Bayesian regression model classifier for breast cancer based on image features of digitized mammograms. Evaluate the model performance using honest leave-one-out cross-validation (LOOCV) with the ROC area as the performance metric. Calculate the Bayesian posterior classification probability intervals to provide an honest assessment of the uncertainties of the predictive classifications. (Months 1-12)

This task has been completed and has resulted in publications (see #1, #2, and #3 in Reportable Outcomes). On each digitized mammogram, a 512x512 region of interest (ROI) centered on the centroid of each calcification cluster was extracted. The automated image-processing scheme consisted of the following steps: (1) pre-processing using unsharp masking, (2) segmentation of individual calcifications using a back-propagation artificial neural network (BP-ANN) classifier, and (3) cluster classification using another BP-ANN classifier to reduce the number of false positive clusters. For each cluster, the algorithm calculated 22 image-processing features, consisting mostly of shape features for the calcifications and calcification clusters and of texture features for ROIs centered on the clusters.

Once the features had been extracted from the mammogram, they were used to distinguish benign from malignant calcification lesions by classification models. In addition to Bayesian probit regression models, for comparison we also applied two well-established CADx classifiers, linear discriminant analysis (LDA), artificial neural network (ANN). We also applied two

variants of a novel classifier, decision fusion: decision fusion to maximize the area under the ROC curve (DF-A), and to maximize the high-sensitivity region (TPF ≥ 0.90) partial area (DF-P). Decision fusion was a novel classification method (See #1 in Reportable Outcomes). Figure 1a shows the ROC curve for the Bayesian probit regression, and Figure 1b shows the set of ROC curves for the classifiers' performances under 100-fold cross validation were AUC = 0.73 for Bayesian probit regression, 0.68 ± 0.01 for LDA, 0.76 ± 0.01 for ANN, 0.85 ± 0.01 for DF-A, and 0.82 ± 0.01 for DF-P. Decision fusion significantly outperformed the other classifiers ($p < 0.001$).

This result was published in *Medical Phyics*, the premiere peer-reviewed journal in the field of Medical Physics, please see Appendix #1 for the reprinted publication.

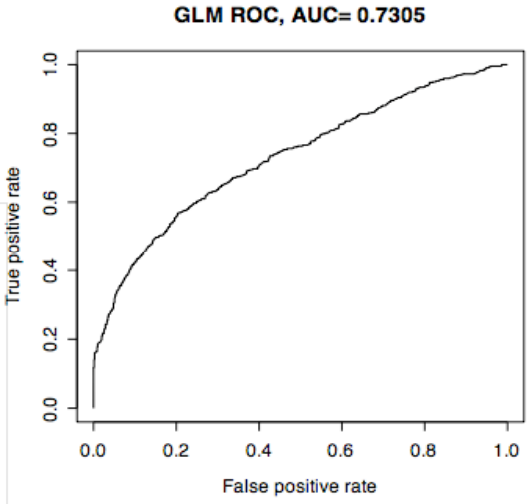


Figure 1a: Bayesian probit regression

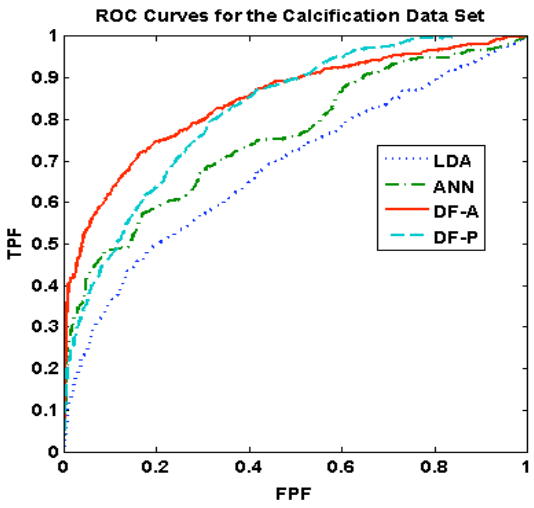


Figure 1b: LDA, ANN, and decision fusion

Task 2. Build a Bayesian regression model classifier for breast cancer based patient age and BI-RADS features from radiologists. Evaluate the model performance and classification uncertainties as in Aim 1. (Months 13-16)

This task has already been completed and has resulted in publications (see #1 and #2 in Reportable Outcomes). The mammographic findings for each case in our database have been interpreted by dedicated breast imaging radiologists using the Breast Imaging Reporting and Data System (BI-RADS) lexicon from the American College of Radiology.[8] The BI-RADS lexicon provides categorical descriptions (findings) for each mammographic feature.

While the original research proposal focused only on microcalcification lesions, we have responded to one of the proposal reviewers and have extended the research project to include masses as well. Including masses will lend additional clinical relevance to this project. Currently, the radiologist-interpreted BI-RADS features are available only for mass cases.

All of the classifiers were able to distinguish benign from malignant lesions well. The classifiers' performances under 100-fold cross validation were AUC = 0.94 for Bayesian probit regression, 0.93 ± 0.01 for LDA, 0.93 ± 0.01for ANN, 0.94 ± 0.01for DF-A, and 0.93 ± 0.01 for DF-P. Decision fusion had a slight performance gain over the ANN and LDA ($p = 0.02$), but was comparable to Bayesian probit regression. The ROC curves of these classifiers are shown in Figures 2a and 2b.

This result was published in *Radiology*, the premiere peer-reviewed journal in the field of Radiology, please see Appendix #2 for the reprinted publication.

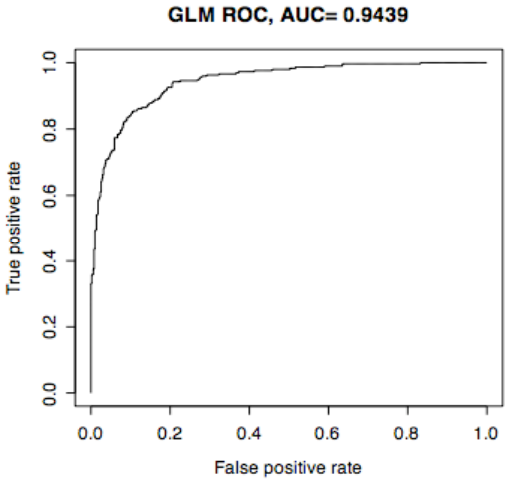


Figure 2a: Bayesian probit regression

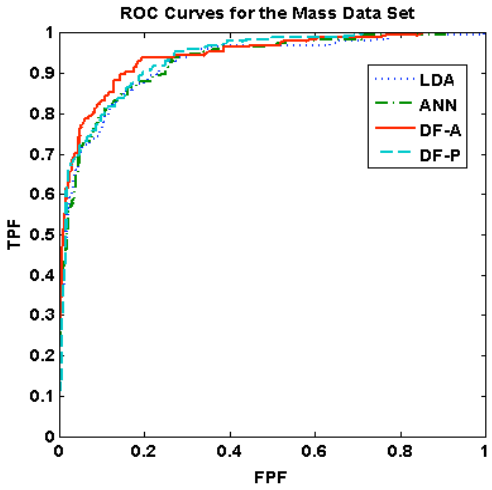


Figure 2b: LDA, ANN, and decision fusion

Task 3. Build a Bayesian regression model classifier for breast cancer based on proteomic profiles of blood serum samples. Evaluate the model performance and classification uncertainties as in Aim 1. (Months 16-28)

We have completed this task and have submitted our results for publication (see #6 in Reportable Outcomes).

This study enrolled 165 premenopausal women undergoing diagnostic biopsy at Duke Univeristy Medical Center for breast cancer between 1999-2005. Before cytoreductive surgery, women were consented for the study and blood was obtained. Serum, plasma, and white blood cells were aliquoted and cryogenically stored. Three sets were constructed from these samples: 1) 48 benign subjects and 2) 49 subjects with invasive breast cancers greater than 1.5 cm, and 3) 68 healthy subjects as controls.

While the original research proposal included proteomic data from mass spectrometry spectra, these spectra were found to be too noisy for the purposes of classifying malignant from benign lesions. We used the much more specific Enzyme-Linked ImmunoSorbent Assay (ELISA) protocol to extract information about blood serum proteins. Sera were assayed for 98 different biomarkers using the Luminex platform and reagents (see #6 in Reportable Outcomes).

To model explicitly the uncertainty due to model selection and for more robust prediction, we used Bayesian model averaging methods. These methods were compared with more traditional classifiers. Figure 3a shows the selected models for normal vs. cancer for iterated Bayesian model averaging of linear models. Models are ordered by selection frequency, with the best, most frequently selected models on the left and the weakest, rarest chosen on the right. Coefficients with positive values are shown in red and negative values in blue. Strong, frequently selected features appear as solid horizontal stripes, such as for MIF, patient age, MMP-9, and MPO. Figures 3b,c,and d show the ROC curves for the three binary classification tasks. The proteins allowed the models to detection lesions moderately well (AUC = 0.82 for normal tissue vs. malignant lesions and for normal tissue vs. benign lesions). However, the benign and malignant lesions had nearly identical serum protein compositions, resulting in very poor classification performance (AUC = 0.55). The selected proteins likely play a role in the

inflammatory response to a lesion, whether benign or malignant, rather than in a role specific for cancer.

This result was submitted to *Bioinformatics*, the premiere peer-reviewed journal in the field of Bioinformaics, please see Appendix #6 for the reprinted publication.

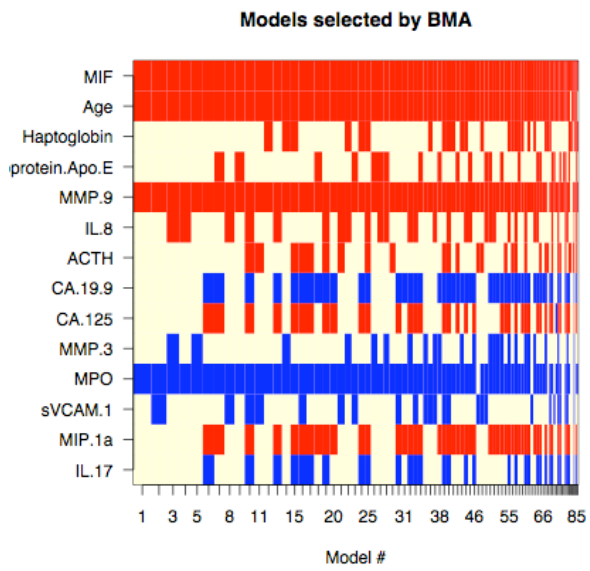


Figure 3a: Selected models for iBMA of linear models for normal tissue vs. malignant lesions

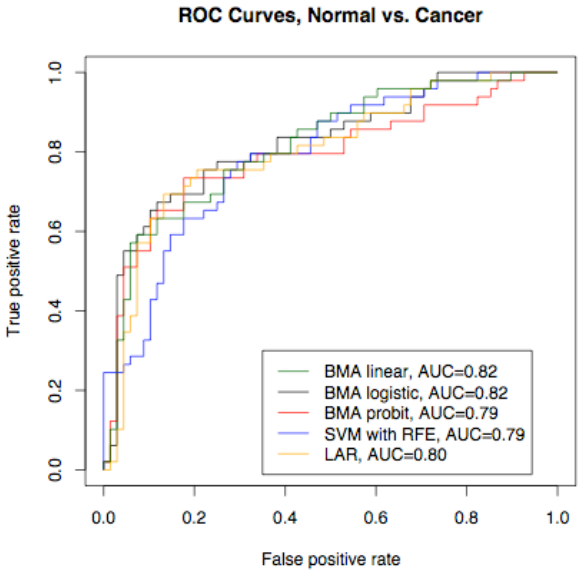


Figure 3b: ROC curves for normal tissue vs. malignant lesions

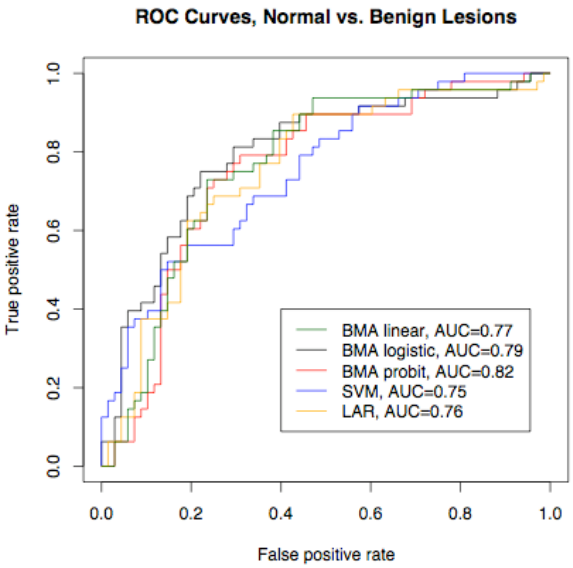


Figure 3c: ROC curves for normal vs. benign lesions

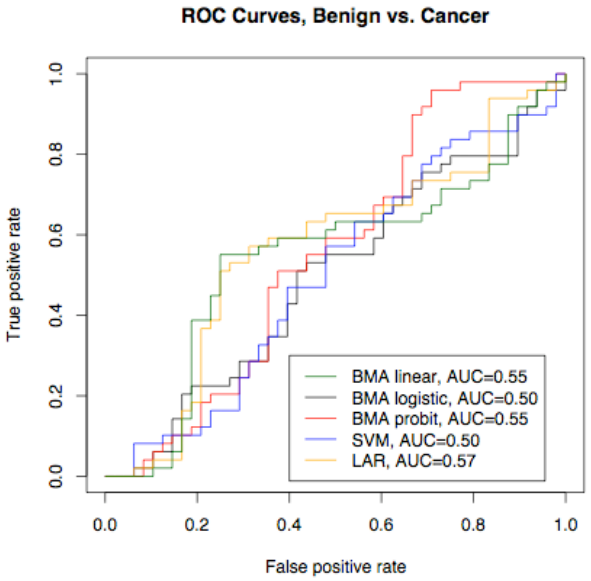


Figure 3d: ROC curves for benign vs. malignant lesions

Task 4. Combine the outputs of the three Bayesian regression models into one ensemble classifier for breast cancer diagnosis prediction. Evaluate the model performance using the ROC area as the performance metric. (Months 28-36)

Between the independently collected radiology data set (737 subjects) and the proteomics data set (165 subjects) there were unfortunately only 12 overlapping subjects. Such a small sample size did not allow for the development of robust predictive models to integrate the data sources. To allow for computational modeling, further data collection efforts were scheduled for the third budget year. However, we are relinquishing the third year because the predoctoral fellowship recipient, Jonathan Jesneck, has just received his PhD and graduated. Other graduate students in the lab, however, intend to continue the research project and to combine the data sources for future studies.

KEY RESEARCH ACCOMPLISHMENTS

- Developed a decision fusion model to combine various information sources
- Classified the mammogram and BI-RADS data sets using the following classification models: Bayesian probit regression, linear discriminant analysis, artificial neural network, and decision fusion
- Established an internal collaboration as a data source for the proteomics data set, and initiated preliminary analysis of that data set.
- Classified the proteomics data set using the following classification models: iterated Bayesian model averaging of linear, logistics, and probit models; support vector machine, and least-angle regression.
- Jonathan Jesneck received a graduate Certificate in Computational Biology and Bioinformatics.
- Jonathan Jesneck received a MS in Statistics and Decision Sciences.
- Jonathan Jesneck received a PhD in Biomedical Engineering.

CONCLUSIONS

The current work focuses on combining breast imaging and proteomics information for breast cancer diagnosis. This study is structured in two stages: (1) build classification models on each of the individual data sources, and (2) combine the models into one ensemble classifier.

One significant research outcome was the development of a decision fusion classification algorithm. Decision fusion has the benefit of being robust in very noisy data sets, such as the calcification and proteomics data sets. On the more challenging calcification data set, decision fusion outperformed the other classifiers by achieving AUC = 0.85 ± 0.01. On the BI-RADS data set, all classifiers performed well, with decision fusion still performing the best with AUC = 0.94 ± 0.01.

The proteomics study showed that serum proteins can detect the presence of a lesion reasonably well (AUC = 0.82), but they did not distinguish benign from malignant lesions (AUC = 0.55). The selected proteins showed evidence of secondary effects, such as inflammatory response, rather than acting as a biomarker specific for cancer.

Classifiers worked well for the separate radiology data set and the proteomics data set, but we were prevented from building predictive models combining the information due to the very small number of overlaps between the two data sets. Future data collection efforts for these data sets will be coordinated as to maximize overlapping cases.

REPORTABLE OUTCOMES

The following publications are attached as appendices 1-4 with the same numbers. The names of the fellow (Jesneck) and mentor (Lo) are boldfaced for emphasis.

1 **Jesneck JL**, Nolte LW, Baker JA, Floyd CE, **Lo JY**, "An optimized approach to decision fusion of heterogeneous data for Breast Cancer Diagnosis*," Medical Physics*, 2006, 33(8):2945-54.

2 **Jesneck JL**, Nolte LW, Baker JA, **Lo JY**, "The effect of data set size on computer-aided diagnosis of breast cancer: Comparing decision fusion to a linear discriminant," in SPIE medical Imaging 2006: Image Processing (2006) 6146, 614616

4 **Jesneck JL**, Nolte LW, **Lo JY**, "An optimized decision-fusion algorithm for classification of heterogeneous breast cancer data," *Proc. SPIE*, accepted for publication in February 2006

5 **Jesneck JL**, **Lo JY**, Baker JA, "A computer aid for diagnosis of breast mass lesions using both mammographic and sonographic BI-RADS descriptors," *Radiology*, 2007 Jun 11; [Epub ahead of print]

6 **Jesneck JL**, Nolte LW, Tourassi GD, **Lo JY**, "A Bayesian method to estimate the minimum sample size for decision fusion," *Medical Decision Making*, submitted

7 Tourassi GD, **Jesneck JL**, Mazurowski M, Habas P, "Stacked Generalization in Computer-Assisted Decision Systems: Empirical Comparison of Data Handling Schemes," *Proc. SPIE*, accepted for publication in March 2007

8 **Jesneck JL**, Mukherjee S, Lokshin AE, Marks JR, Clyde M, Lo JY, "Identifying circulating protein markers for breast cancer detection in premenopausal women," *Bioinformatics*, submitted

REFRENCES

1. S. Shapiro, "Screening: assessment of current studies", *Cancer* 1994; 74:231–238.
2. I.C. Henderson, "Breast cancer", In: Murphy GP, W. Lawrence Jr, R.E. Lenhard, eds. *American Cancer Society textbook of clinical oncology*. Atlanta, Ga: American Cancer Society, 1995; 198–219.
3. A.M. Knutzen, J.J Gisvold, "Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions," *Mayo Clin Proc* 1993; 68: 454-460.
4. D.B. Kopans, "The Positive Predictive Value of Mammography," *AJR* 1992; 158:521-526.
5. A. S. Hong, E. L. Rosen, M. S. Soo et al., "BI-RADS for Sonography: Positive and Negative Predictive Values of Sonographic Features." *AJR* 184 (4), 1260 (2005).
6. E.F. Petricoin, C.P. Paweletz, and L.A. Liotta, "Clinical Applications of Proteomics: Proteomic Pattern Diagnostics", *Journal of Mammary Gland Biology and Neoplasia*, Vol. 7, No. 4, October 2002, p.433-440.
7. J.Y. Lo, M. Gavrielides, M.K. Markey, J.L. Jesneck, "Computer-aided classification of breast microcalcification clusters: merging of features from image processing and radiologists", *Proc. SPIE* Vol. 5032, p. 882-889, Medical Imaging 2003: Image Processing; Milan Sonka, J. Michael Fitzpatrick; Eds.

8. BI-RADS. American College of Radiology. *American College of Radiology Breast Imaging - Reporting and Data System (BI-RADS) 3rd ed.*, 1998.

APPENDICES

Six publications are attached, see "Reportable Outcomes" above for the list.

Appendix

Optimized approach to decision fusion of heterogeneous data for breast cancer diagnosis

Jonathan L. Jesneck[a)]
Department of Biomedical Engineering, Duke University, Durham, North Carolina 27705 and Duke Advanced Imaging Labs, Duke University, Durham, NC 27705

Loren W. Nolte
Department of Biomedical Engineering, Duke University, Durham, North Carolina 27705 and Department of Electrical and Computer Engineering, Duke University, Durham, NC 27705

Jay A. Baker
Duke Advanced Imaging Labs, Department of Radiology, Duke University, Durham, North Carolina 27705

Carey E. Floyd and Joseph Y. Lo
Department of Biomedical Engineering, Duke University, Durham, North Carolina 27705 and Duke Advanced Imaging Labs, Duke University, Durham, NC 27705 and Medical Physics Graduate Program, Duke University, Durham, NC 27705

(Received 1 January 2006; revised 5 April 2006; accepted for publication 4 May 2006; published 26 July 2006)

As more diagnostic testing options become available to physicians, it becomes more difficult to combine various types of medical information together in order to optimize the overall diagnosis. To improve diagnostic performance, here we introduce an approach to optimize a decision-fusion technique to combine heterogeneous information, such as from different modalities, feature categories, or institutions. For classifier comparison we used two performance metrics: The receiving operator characteristic (ROC) area under the curve [area under the ROC curve (AUC)] and the normalized partial area under the curve (pAUC). This study used four classifiers: Linear discriminant analysis (LDA), artificial neural network (ANN), and two variants of our decision-fusion technique, AUC-optimized (DF-A) and pAUC-optimized (DF-P) decision fusion. We applied each of these classifiers with 100-fold cross-validation to two heterogeneous breast cancer data sets: One of mass lesion features and a much more challenging one of microcalcification lesion features. For the calcification data set, DF-A outperformed the other classifiers in terms of AUC ($p<0.02$) and achieved AUC=0.85±0.01. The DF-P surpassed the other classifiers in terms of pAUC ($p<0.01$) and reached pAUC=0.38±0.02. For the mass data set, DF-A outperformed both the ANN and the LDA ($p<0.04$) and achieved AUC=0.94±0.01. Although for this data set there were no statistically significant differences among the classifiers' pAUC values (pAUC=0.57±0.07 to 0.67±0.05, $p>0.10$), the DF-P did significantly improve specificity versus the LDA at both 98% and 100% sensitivity ($p<0.04$). In conclusion, decision fusion directly optimized clinically significant performance measures, such as AUC and pAUC, and sometimes outperformed two well-known machine-learning techniques when applied to two different breast cancer data sets. [DOI: 10.1118/1.2208934]

Key words: decision fusion, heterogeneous data, receiver operating characteristic (ROC) curve, area under the curve (AUC), partial area under the curve (pAUC), classification, machine learning, breast cancer

I. INTRODUCTION

Breast cancer accounts for one-third of all cancer diagnoses among American women, has the second highest mortality rate of all cancer deaths in women,[1] and is expected to account for 15% of all cancer deaths in 2005.[2] Early diagnosis and treatment can significantly improve the chance of survival for breast cancer patients.[3] Currently, mammography is the preferred screening method for breast cancer. However, high false positive rates reduce the effectiveness of screening mammography, as several studies have shown that only 13–29% of suspicious masses are determined to be malignant.[4–6] Unnecessary surgical biopsies are expensive, cause patient anxiety, alter cosmetic appearance, and can distort future mammograms.[7]

Commercial products for computer-aided detection (CAD) have shown promise for improving sensitivity in large clinical trials. Most studies to date have shown CAD to boost radiologists lesion detection sensitivity.[8–11] To date, however, there are no commercial systems to improve specificity for breast cancer screening. To fill this need to improve the sensitivity of mammography, computer-aided diagnosis (CADx) has emerged as a promising clinical aid.[12]

There has been considerable CAD and CADx research based upon a rich variety of modalities and sources of medical information, such as: digitized screen-film mammograms,[13–17] full-field digital mammograms,[18] sonograms,[19–21] magnetic resonance imaging (MRI) images,[22] and gene expression profiles.[23] Current clinically implemented CADx programs tend to use only one information source, although multimodality CADx programs[24] are beginning to emerge. Moreover, most CADx research has been performed using relatively homogeneous data sets collected at one institution, acquired using one type of digitizer or digital detector, or using features drawn from one source such as human-interpreted findings versus computer-extracted features. Increasingly however, there is a trend toward boosting diagnostic performance by combining data from many different sources to create heterogeneous data. We defined heterogeneous data as comprising multiple, distinct groups. Specifically, for this study, we considered as heterogeneous any of the following data set characteristics: Multiple imaging modalities, multiple types of mammogram film digitizers, data collected from multiple institutions, and various types of features extracted from the same image, especially computer-extracted and human-extracted features. Combining heterogeneous data types for classification is a difficult machine-learning problem, but one that has shown promise in bioinformatics applications.[25–27]

To meet the challenge of combining heterogeneous data types, we turned to a decision-fusion method that operates by the following two steps: (1) Classifiers use feature subsets to generate initial binary decisions, and (2) these binary decisions are then optimally combined by using decision-fusion theory. Decision fusion offers the following advantages: It handles heterogeneous data sources well, reduces the problem dimensionality, is easily interpretable, and is easy to use in a clinical setting. Decision fusion has effectively combined heterogeneous data in many diverse classification tasks, such as detecting land mines using multiple sensors,[28] identifying persons using multiple biometrics,[29] and CADx of endoscopic images using multiple sets of medical features.[30]

The purpose of this study was to optimize a decision-fusion approach for classifying heterogeneous breast cancer data. We compared this decision-fusion approach to a linear discriminant and an artificial neural network (ANN), which are well-studied techniques that have frequently been applied to breast cancer CADx.[13,31–33] This study evaluates these classification algorithms on two breast cancer data sets using two different clinically relevant performance metrics.

II. METHODS

A. Data

For this study, we chose two different breast cancer data sets, which differed considerably in the type and number of patient cases as well as the type and number of medical information features describing those cases.

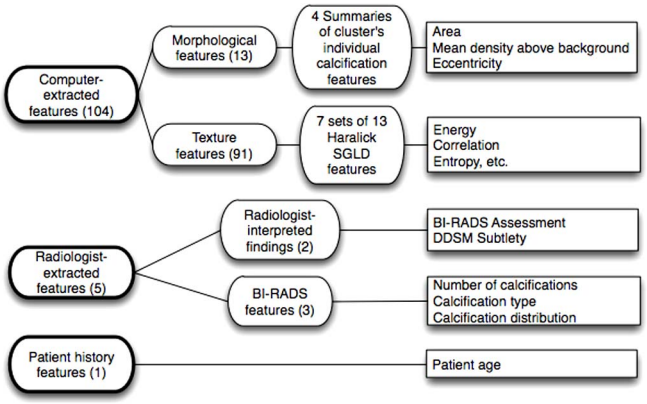


FIG. 1. Feature group structure for calcification Data Set C (calcification lesions). The features of the calcification data set consisted of three main groups: Computer-extracted features, radiologist-extracted features, and patient history features. The computer-extracted features were morphological and shape features of the automatically detected and segmented microcalcification clusters within the digitized mammogram images. The radiologist-extracted features comprised both radiologist-interpreted findings and BI-RADS features. This data set consisted of 512×512 pixel ROIs of all 1508 calcification lesions in the DDSM. This data set had many heterogenic characteristics, such as that it was collected at four different institutions, scanned on four digitizers with different noise characteristics, and included both human-extracted and computer-extracted features, such as shape and texture features.

1. Microcalcification lesions

Data set C consisted of all 1508 mammogram microcalcification lesions from the Digital Database for Screening Mammography (DDSM).[34] The outcomes were verified by histological diagnosis and followup for certain benign cases, yielding 811 benign and 697 malignant calcification lesions. Figure 1 shows the feature group structure of this data set. The feature groups were 13 computer-extracted calcification cluster morphological features, 91 computer-extracted texture features of the lesion background anatomy, 2 radiologist-interpreted findings, 3 radiologist-extracted features from the Breast Imaging Reporting and Data System (BI-RADS™, American College of Radiology, Reston, VA) (Ref. 35) and patient age. In total, data set C had 110 features and a sample-to-feature ratio of approximately 14:1. Each mammogram was digitized with one of four digitizers: A DBA M2100 ImageClear at a resolution of 42 microns, a Howtek 960 at 43.5 microns, a Howtek MultiRad850 at 43.5 microns, or a Lumisys 200 Laser at 50 microns. To study this large heterogeneous data set, no attempt was made to restrict cases only to a single digitizer, as was common in most previous studies. Moreover, no standardization step was applied to the images to correct for the differences in noise, resolution, and other physical characteristics from the various digitizers. We used a 512×512 pixel region of interest (ROI) centered on the centroid of each lesion (using lesion outlines drawn by the DDSM radiologists) for image processing and for generating the computer-extracted features. We extracted morphological and texture (spatial gray level dependence matrix) features, which were shown to be useful in a previous study of CADx by Chan *et al.*[31]

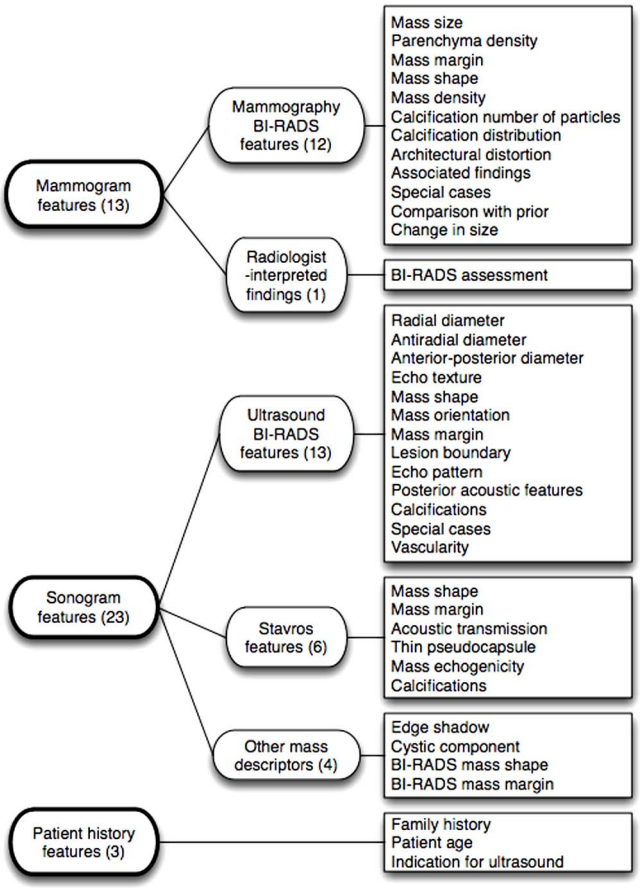


FIG. 2. Feature group structure for mass Data set M (mass lesions). The features of the mass data set consisted of mammogram features, sonogram features, and patient history features. The mammogram features comprised both BI-RADS features and radiologist-interpreted findings. The sonogram features consisted of ultrasound BI-RADS features, Stavros features, and other ultrasound mass descriptors. All image features were radiologist-extracted features. The mass data set was heterogeneous in including both mammogram and sonogram views of the breast. Both mammogram and sonogram feature sets were as well as including patient history features.

This data set had many heterogenic characteristics, such as that it was collected at four different institutions, scanned on four types of digitizers with different physical characteristics, and included both human-extracted and computer-extracted features, such as shape and texture features.

2. Mass lesions

Data set M consisted of 568 breast mass cases that were collected in the Radiology Department of Duke University Health System between 1999 and 2001. These cases were an extension of the data set described in detail in our previous studies.[36,37] Definitive histopathologic diagnosis from biopsy was used to determine outcome, yielding 370 benign and 198 malignant mass lesions. Figure 2 shows the feature group structure of this data set. Dedicated breast radiologists recorded all features.

The mass data set was heterogeneous because it was comprised of 3 distinct types of data: 13 mammogram features, 23 sonogram features in turn drawn from 3 different lexicons

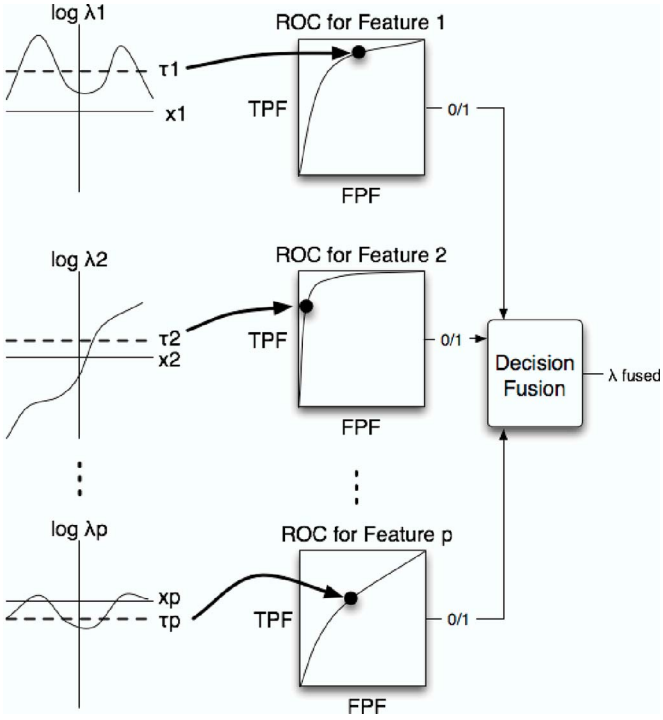


FIG. 3. The role of likelihood-ratio thresholds for decision fusion. The first column shows plots of the log-likelihood-ratio versus feature value for each feature. The algorithm calculated the likelihood ratio and then thresholded it separately for each feature. The threshold determined the ROC operating point of the likelihood-ratio classifier of a particular feature. Next, the algorithm combined the binary decisions from the feature-level likelihood ratio classifiers using decision fusion theory to produce the likelihood ratio of the fused classifier.

(Ultrasound Bl-RADS, Stavros, and others),[36] as well as 3 patient history features. In total, data set M had 39 features and a sample-to-feature ratio of approximately 15:1.

B. Decision fusion

There is a growing literature in the area of distributed detection. Although there is even some earlier work, several of the early classical references include the work of Tenney and Sandell,[38] who introduced distributed detection using a fixed fusion processor and optimized the local processors. Chair and Varshney[39] fixed the local processors, and optimized the fusion processor. Reibman and Nolte[40] extended these previous studies by simultaneous optimization of the local detectors while deriving the overall optimum fusion design. Dasarathy[41] summarized some of the earlier work.

Decision-fusion theory describes how to combine local binary decisions optimally to determine the presence or absence of a signal in noise.[38–42] The local binary decisions can come from any arbitrary source.

Figure 3 provides a schematic of our decision-fusion method. Our algorithm is a two-stage process, each with a likelihood ratio calculation. The first stage applies a separate likelihood ratio to each feature. These feature-level likelihood ratios are then compared to separate thresholds to generate feature-level decisions. These feature-level decisions are then fused in the second stage by computing the likeli-

hood ratio of the binary decision values. The second stage combines the feature-level decisions into one fused likelihood-ratio value, which can be used as a classification decision variable.

Our technique offers the important advantage that it can reduce the dimensionality of the feature space of the classification problem by assigning a classifier to each feature separately. Considering only one feature at a time greatly reduces the complexity of the problem by avoiding the need to estimate multidimensional probability density functions (PDFs) of the feature space. Accurately estimating such multidimensional PDFs likely requires many more observations than a typical medical data set contains. Other benefits of decision fusion are that it is robust in noisy data,[43] is not overly sensitive to the likelihood ratio threshold values,[42] and can handle missing data values.[44] Our decision-fusion technique can also be tuned to maximize arbitrary performance metrics (as described later in Sec. II C) that may be more clinically relevant, unlike more traditional classification algorithms that minimize mean-squared error.

1. Detection theory approach - likelihood ratio

Although decision fusion combines binary decisions regardless of how those decisions were made, it is still important to choose the right initial classifiers in order to pass as much information to the decision fuser as possible. In our algorithm, we used the likelihood ratio as the initial classifier and applied a threshold to generate the binary decisions on each feature. Previous work has shown the likelihood ratio to be an excellent classifier for breast cancer mass lesion data.[45,46]

According to decision theory, the likelihood ratio is the optimal detector to determine the presence or absence of a signal in noise.[47] For this study, the signal to be detected was the potential malignancy of a breast lesion. The null hypothesis (H_0) was that the signal (malignancy) is not present in the noisy features, while the alternative hypothesis (H_1) was that the signal is present:

$$H_0: X = N,$$

$$H_1: X = S + N. \tag{1}$$

Sources of noise in the features included anatomical noise inherent in the mammogram or sonogram, quantum noise in the acquisition of the mammogram or sonogram, digitization noise and artifacts for data set C, and ambiguities in the mammogram reading process for the radiologist-interpreted findings in both Data sets C and M.

The likelihood ratio is the probability of the features under the malignant case divided by the probability of the features under the benign case:

$$\lambda_{\text{features}}(X) = \frac{P(X|H_1)}{P(X|H_0)}, \tag{2}$$

where $P(X|H_1)$ is the PDF of the observation data X given that the signal is present, and $P(X|H_0)$ is the PDF of the data X given that the signal is not present. The likelihood ratio is optimal under the assumption that the PDFs accurately reflect the true densities. We estimated the one-dimensional PDFs of the features with histograms. We used Scott's rule to determine the optimal histogram bin width,[45]

$$h = 3.5\sigma n^{-1/3}, \tag{3}$$

where h is the bin width, σ is the standard deviation, and n is the number of observations. The interval of two standard deviations around the mean, $[\mu-2\sigma, \mu+2\sigma]$, was then subdivided by the bin width, h. We assigned the values falling outside this interval to the extreme left or right bins. Next, we applied a threshold value, τ, to the likelihood ratio to produce a binary decision about the presence of the signal.

$$u = \begin{cases} 1 & \text{if } \lambda_{\text{feature}} \geqslant \tau \\ 0 & \text{if } \lambda_{\text{feature}} < \tau. \end{cases} \tag{4}$$

2. Fusing the binary decisions

For the signal-plus-noise hypothesis H_1, the probability of detecting an existing signal is $P(u=1|H_1)=Pd$ and of missing it is $P(u=0|H_1)=1-Pd$. For the noise-only hypothesis H_0, the probability of false detection is $P(u=1|H_0)=Pf$ and of correctly rejecting the missing signal is $P(u=0|H_0)=1-Pf$. Using these probabilities, the likelihood ratio value of a binary decision variable has a simple form, as shown in Eq. (5):

$$\lambda_{\text{decision}}(u) = \frac{P(u|H_1)}{P(u|H_0)} = \begin{cases} \dfrac{Pd}{Pf} & \text{if } u = 1 \\ \dfrac{1-Pd}{1-Pf} & \text{if } u = 0. \end{cases} \tag{5}$$

We can then use the likelihood ratios of the individual local decision variables to calculate the joint likelihood ratio of the set of decision variables. Assuming that the local decision variables are statistically independent, the likelihood ratio of the fused classifier is a product of the likelihood ratios of the individual local decisions.

$$\begin{aligned} \lambda_{\text{fusion}}(u_1, \ldots, u_p) &= \prod_{i=1}^{p} \lambda_{\text{decision}}(u_i) \\ &= \prod_{i=1}^{p} \frac{P(u_i|H_1)}{P(u_i|H_0)} \\ &= \prod_{i=1}^{p} \left(\frac{Pd_i}{Pf_i}\right)^{u_i} \left(\frac{1-Pd_i}{1-Pf_i}\right)^{1-u_i}. \end{aligned} \tag{6}$$

Note that we assume statistical independence of only the local binary decisions, not of the sensitivity, false-positive rate, or even the features on which the local decisions were made.

In our decision-fusion theory approach, we have made the important assumption that all the local decisions are statistically independent. While this appears to be a very strong assumption, using it in decision fusion often does not lower classification performance substantially below the perfor-

mance of the optimal decision fusion processor for correlated decisions. Although we can construct an optimal correlated decision-fusion processor with known decision correlations,[48] it is difficult to estimate the correlation structure of the decisions accurately, especially given many decisions but only few observations. However, even with correlated decisions, the simplifying assumption of independent decisions often does not lower decision fusion performance. Liao *et al.*[42] have shown that, under certain conditions for the case of fusing two correlated decisions, the independent fusion processor exactly matched the performance of the optimal correlated decision fusion processor. Even in many situations when the optimality conditions were not kept, the degradation of the fusion performance was not significant.[42] Another benefit of the independent local decisions assumption is that decision fusion can usually recover from weak signals and correlated features given enough decisions to fuse.[43] Because we have a large number of local decisions by setting a separate local decision for each feature, our algorithm takes advantage of this performance benefit.

C. Classifier evaluation and figures of merit

We used the receiver operating characteristic (ROC) curve to capture the classification performance of our decision-fusion algorithm. Assuming independent local decisions, the PDFs of the decision-fusion likelihood ratio have a similar product form.[42]

$$P(\lambda_{\text{fusion}}|H_1) = \prod_{i=1}^{p} (Pd_i)^{u_i}(1 - Pd_i)^{1-u_i},$$

$$P(\lambda_{\text{fusion}}|H_0) = \prod_{i=1}^{p} (Pf_i)^{u_i}(1 - Pf_i)^{1-u_i}. \quad (7)$$

Using the fusion likelihood ratio value as a classification decision variable, the probabilities of detection and false alarm are calculated as follows:

$$Pd_{\text{fusion}}(\beta) = \sum_{\lambda_{\text{fusion}} \geqslant \beta} P(\lambda = \lambda_{\text{fusion}}|H_1),$$

$$Pf_{\text{fusion}}(\beta) = \sum_{\lambda_{\text{fusion}} \geqslant \beta} P(\lambda = \lambda_{\text{fusion}}|H_0), \quad (8)$$

where β is a threshold on λ_{fusion} that determines the operating point on the ROC curve. By varying the value of the threshold β, these $Pd_{\text{fusion}}(\beta)$ and $Pf_{\text{fusion}}(\beta)$ values trace the entire decision-fusion ROC curve.

One can use the ROC curve to quantify classification performance by calculating summary metrics of the curve. Certain performance metrics have more significance in a clinical setting than others, especially when high sensitivity must be maintained. This study used two clinically interesting summary metrics of the ROC curve: The area under the curve (AUC), and the normalized partial area under the curve (pAUC) above a certain sensitivity value.[49] For this study, we set the sensitivity value true positive fraction (TPF) $=0.90$ for pAUC to reflect that diagnosing breast cancer at high sensitivities is clinically imperative. We used the nonparametric bootstrap method[50] to measure the means and variances of the AUC and pAUC values as well as to compare metrics from two models for statistical significance.

D. Genetic algorithm search for the optimal threshold set

The selection of the likelihood-ratio threshold values is important to maximize performance of the fused classifier. Threshold values very far from the best values often lowered the fused classifier's performance to near chance levels. A genetic algorithm searched over the likelihood-ratio threshold values for each feature to select a threshold set that maximized the desired performance metric or figure of merit (FOM),

$$\tau_{\text{optimal}} = \text{argmax FOM}[\lambda_{\text{fusion}}(u;\tau)], \quad (9)$$

where the FOM is either AUC or pAUC, u is the set of local decisions, and τ is the set of feature-level likelihood-ratio thresholds. The fitness function of the genetic algorithm was set to the FOM in order to maximize the FOM value. We optimized for cross-validation performance the following genetic algorithm parameters: The number of generations, population size, and rates of selection, crossover, and mutation.

E. Decision fusion with cross-validation

We used k-fold cross-validation ($k=100$) to estimate the ability of the classifiers to generalize on our data sets. For each fold, a new model was developed, i.e., the likelihood ratio was formed on the $k-1$ subsets (99% of cases) used as training samples, and the genetic algorithm searched over the thresholds to maximize the performance metric on these training samples. Once the best thresholds had been found on the training set, they were then used to evaluate the algorithm on the one subset (1% of cases) withheld for validation. The resulting local decisions were then combined into the fused validation likelihood ratio $\lambda_{\text{test,fusion}}$, as in Eq. (6). The process was then repeated k times by withholding a different subset for validation, such that all cases are used for training and validation while simultaneously ensuring independence between those subsets.

Compiling all $\lambda_{\text{test,fusion}}$ values at the end of the cross-validation computations created a distribution of $\lambda_{\text{test,fusion}}(X)$ of the test cases. We constructed an ROC curve from the $\lambda_{\text{test,fusion}}(X)$ values, as in Eq. (8), in order to measure the classification performance of the decision-fusion classifier with k-fold cross-validation.

F. Using decision fusion in a diagnostic setting

Once the model has been fully trained and validated, it can similarly be applied to new cases by setting all of the existing data to be the training data and applying the new clinical case as a new validation case. The decision-fusion

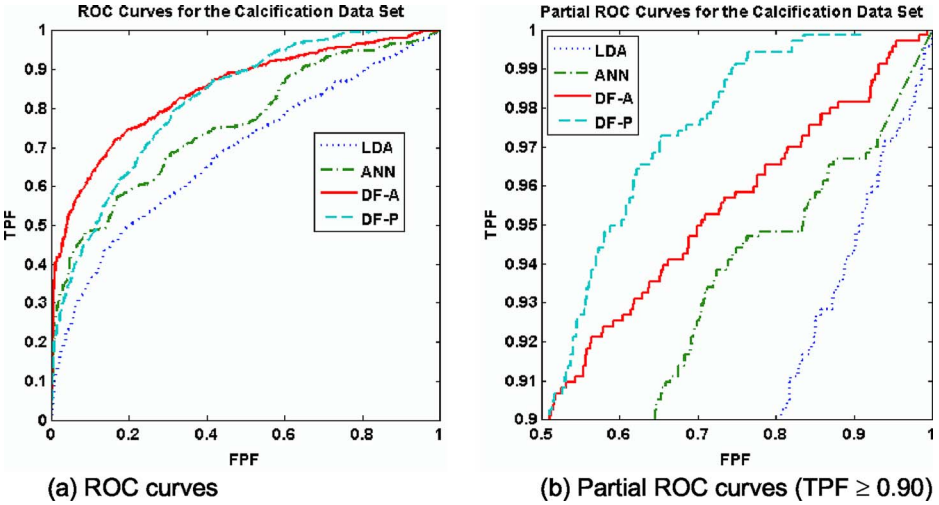


FIG. 4. ROC curves for Data set C (calcification lesions). The classifiers' ROC curves for 100-fold cross-validation are shown. Figure 4(a) shows the full ROC curves, while Figure 4(b) shows only the high-sensitivity region (TPF$\geqslant$0.90). For the calcification data set, the four classifiers yielded differing classification performance under 100-fold cross-validation. Both decision-fusion curves lay significantly above the LDA and ANN curves, both in terms of AUC and pAUC. As expected, the decision-fusion classifiers achieved the highest scores of all the classifiers for their target performance metrics; DF-A attained the greatest AUC, whereas DF-P attained the greatest pAUC. The DF-P curve surpassed the DF-A curve and dominated the other curves above the line TPF=0.90. In order to gain high-sensitivity performance, DF-P sacrificed performance in the less clinically relevant range of TPF$<$0.90.

algorithm would recommend to the physician either a biopsy with a malignant classification or short-term follow-up with a very likely benign classification.

G. Other classifiers: Artificial neural network and linear discriminant

We compared the classification performance of the decision fusion against both an ANN and Fisher's linear discriminant analysis (LDA), which are well-understood algorithms and are popular breast cancer CADx research tools.

For the ANN, we used a fully connected, feed-forward error backpropagation network with a hidden layer of five nodes, implemented using the nnet package (version 7.2-20) for statistical software (version 1.12, the R Project for Statistical Computing). For the LDA, we used the Statistics Toolbox (version 5.1) of MATLAB® (Release 14, Service Pack 2, Mathworks Inc, Natick, MA). Both models were carefully verified against custom software previously developed within our group. We implemented our decision-fusion algorithm in MATLAB, relying specifically on the Genetic Algorithm and Direct Search Toolbox (version 2) to find the best thresholds for the likelihood ratio values.

III. RESULTS

A. Classifier performance on data set C (calcification lesions)

Figure 4 shows the validation ROC curves for the calcification data. Table I lists the classification performances of the four classifiers, while Tables II and III list the two-tailed p values for the pairwise comparisons by AUC and pAUC, respectively. The AUC-optimized decision fusion (DF-A) showed the best overall performance, with AUC $=0.85\pm0.01$, and the pAUC-optimized decision fusion (DF-P) was slightly worse with AUC$=0.82\pm0.01$. Both decision-fusion ROC curves were well above those of the LDA and ANN, both in terms of AUC ($p<0.0001$) and pAUC ($p<0.02$). None of the features were particularly strong by themselves; we ran an LDA on each feature separately, yielding on average AUC$=0.53\pm0.03$, with a maximum of AUC$=0.66$ for the best feature.

The DF-P curve (pAUC$=0.38\pm0.02$) crossed the DF-A curve (pAUC$=0.28\pm0.03$) at the line TPF $=$ 0.9. In order to

TABLE I. Classifier performance on Data set C (calcification lesions). The table shows the AUC and pAUC values for the ROC curves of the four classifiers under 100-fold cross-validation. The performance values exhibited a wide range. The DF-A scored the best for AUC, while DF-P scored highest for pAUC, as expected. The decision-fusion curves soundly outperformed both the ANN and LDA in terms of pAUC.

Classifier	AUC	pAUC
DF-A	0.85±0.01	0.28±0.03
DF-P	0.82±0.01	0.38±0.02
ANN	0.76±0.01	0.14±0.02
LDA	0.68±0.01	0.09±0.06

TABLE II. P values for AUC comparisons for Data set C (calcification lesions). The confusion matrix shows the p values for the pairwise comparisons of the classifiers' AUC values. All pairwise comparisons were statistically significant.

	DF-A	DF-P	ANN	LDA
DF-A		0.018	$<$0.0001	$<$0.0001
DF-P			0.0001	$<$0.0001
ANN				$<$0.0001
LDA				

TABLE III. *P* values for pAUC comparisons for Data set C (calcification lesions). The confusion matrix shows the *p* values for the pairwise comparisons of the classifiers' pAUC values. All pairwise comparisons were statistically significant.

	DF-A	DF-P	ANN	LDA
DF-A		0.0084	0.018	<0.0001
DF-P			0.0001	<0.0001
ANN				0.016
LDA				

gain high-sensitivity performance, DF-P sacrificed performance in the less clinically relevant range of TPF$<$0.9. The DF-A beat the DF-P in terms of AUC (p=0.018) but lost in pAUC ($p<0.01$). Both decision-fusion classifiers greatly outperformed the both the ANN (pAUC=0.14±0.02) and LDA (pAUC=0.09±.06) in terms of pAUC.

B. Classifier performance on data set M (mass lesions)

Figure 5 shows the validation ROC curves of the classifiers for the mass data set. Table IV lists the classification performances of the four classifiers, whereas Tables V and VI list the p values for the pairwise comparisons by AUC and pAUC, respectively. For this data set, all the classifiers had higher but very similar performance, with AUC ranging from 0.93±0.01 (LDA) to 0.94±0.01 (DF-A). With the exception of DF-P (p=0.50), the DF-A nonetheless significantly outperformed both the LDA (p=0.021) and the ANN (p=0.038) in terms of AUC. The LDA, ANN, and DF-P curves were all very similar, for both AUC ($p>0.10$) and pAUC ($p>0.10$). Figure 5(b) shows the ROC curves in the high sensitivity region above the line TPF = 0.90. The classifiers pAUC values ranged narrowly from 0.57±0.07 (ANN) to 0.67±0.05 (DF-P), all close enough to show no statistically significant differences ($p>0.10$). However, the DF-P did have a higher specificity than the LDA at both 98% sensitivity (0.37±0.10 vs. 0.13±0.13, p=0.04) and at 100% sensitivity(0.34±0.08 vs. 0.09±0.12, p=0.03). The DF-P curve passed the DF-A curve approximately at the line TPF = 0.90 and yielded a slightly higher pAUC (0.67±0.05 versus 0.63±0.07), although this improvement was not statistically significant (p=0.48).

IV. DISCUSSION

The multitude of medical data becoming available to physicians presents the problem of how best to integrate the information for diagnostic performance. Despite recent availability of this information, current CADx programs for breast cancer tend to use only one type of data, usually digitized mammogram films. Because many clinical tests provide complementary information about a disease state, it is important to develop a CADx system that incorporates data from disparate sources. However, combining disparate data types together for classification is a difficult machine-learning problem. This study used the likelihood-ratio detector and decision-fusion classifier to detect the presence of a malignancy (a signal) within medical data (noisy features). We also compared the performance of this classifier to two popular classifiers in the CADx literature, LDA and ANN, and we measured the diagnostic performance with two classification metrics, ROC AUC and pAUC. Finally, we performed these studies using two very different data sets in order to assess performance differences due to the data set itself.

Data set C (calcification lesions) had a stronger nonlinear component, indicated by the fact that the ANN AUC was much greater than the LDA AUC. The robustness of the decision-fusion algorithm is evident in its good performance on this weaker, nonlinear, and noisy data set. Decision fusion significantly outperformed the ANN and LDA on the calcification data set for both performance metrics. Figure 4 and Table I show that the biggest performance gain is in the pAUC metric, for which decision fusion doubled the performance of the other classifiers.

On Data set M (mass lesions), all four classifiers seemed to be saturated at a high level of performance in terms of both AUC and pAUC, as shown in Fig. 5 and Table IV. Performances were largely equivalent across all models, ex-

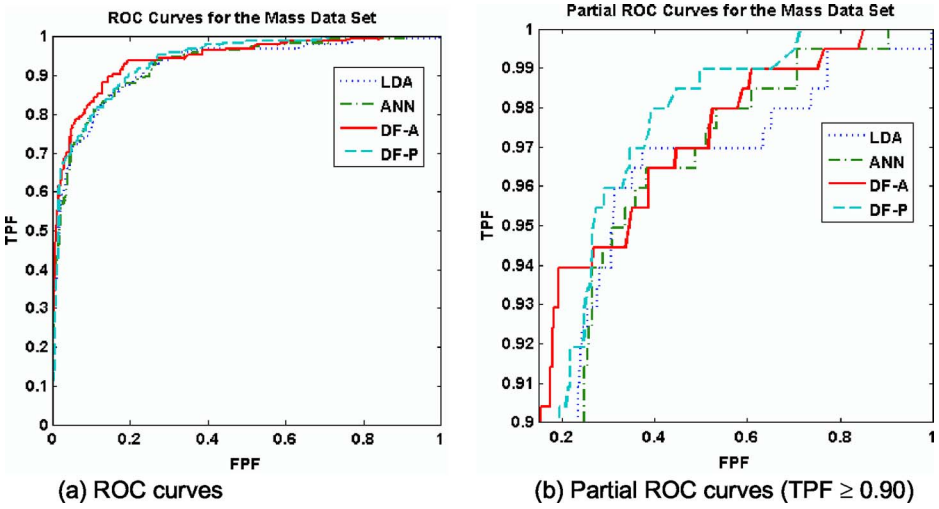


FIG. 5. ROC curves for Data set M (mass lesions). For the mass data set, all classifiers had high levels of classification performance. The DF-A and DF-P achieved the highest AUC and pAUC, respectively. In terms of AUC, the DF-A outperformed both the ANN and LDA (p=0.038 and 0.021, respectively). In (b), the DF-P curve had slightly more partial area than the other curves. Despite having statistically equivalent partial areas, the DF-P had a greater specificity than the LDA at high sensitivities TPF=0.98 (p=0.03).

TABLE IV. Classifier performance on Data set M (mass lesions). The table shows the AUC and pAUC values for the ROC curves of the four classifiers under 100-fold cross-validation. All four classifiers performed very similarly on this data set. The DF-A scored the best for AUC, whereas the DF-P scored highest for pAUC, although both were still within one standard deviation of each of the other classifiers' performances.

Classifier	AUC	pAUC
DF-A	0.94±0.01	0.63±0.07
DF-P	0.93±0.01	0.67±0.05
ANN	0.93±0.01	0.57±0.07
LDA	0.93±0.01	0.59±0.06

cept for two trends. In terms of AUC, the DF-A outperformed both the ANN and the LDA ($p=0.038$ and 0.021, respectively). Although on this data set decision fusion offered only relatively modest gains in pAUC, it did achieve a significantly better specificity than the LDA at several of the highest sensitivities of the ROC curve ($p<0.05$).

This decision-fusion algorithm has many potential benefits over more traditional classification algorithms. Decision fusion can be optimized for any desired performance metric by incorporating the metric into the fitness function of the genetic algorithm for its search over the likelihood-ratio thresholds. This advantage has important clinical implications, as both the physician and the CADx algorithm are constrained to operate at high sensitivity. The performance metric can emphasize good performance at high sensitivities and deemphasize performance at clinically unacceptable low sensitivities. Therefore, we expect the DF-A curve to maximize AUC and the DF-P curve to maximize pAUC. The DF-P curve should fall under the DF-A curve for low FPF values but should cross the DF-A curve at the line TPF $=0.90$ to capture a greater pAUC value. Figures 4 and 5 show evidence that the DF-P did optimize pAUC. The DF-P ROC curves crossed the DF-A curves at the line TPF$=0.90$, and do in fact have a larger pAUC value than the DF-A curves. Another advantage is that decision fusion is robust and can recover from noisy weak features. The likelihood-ratio classifier passes information about the strength or weakness of a feature to the decision fuser, which adjusts the influence given to that feature. This feature-strength information is the ROC operating point (sensitivity and specificity) determined by the likelihood-ratio threshold that was found by the genetic algorithm search. Figure 3 shows a schematic

TABLE V. *P* values for AUC comparisons for Data set M (mass lesions). The confusion matrix shows the *p* values for the pairwise comparisons of the classifiers' AUC values. The DF-A outperformed the ANN and LDA. Among the DF-P, ANN, and LDA, there were no statistically significant pAUC differences.

	DF-A	DF-P	ANN	LDA
DF-A		0.50	0.038	0.021
DF-P			0.20	0.17
ANN				0.53
LDA				

TABLE VI. *P* values for pAUC comparisons for Data set M (mass lesions). The confusion matrix shows the *p* values for the pairwise comparisons of the classifiers' pAUC values. None of the pAUC comparisons were statistically significant. Although pAUC scores were similar, the DF-P did have a higher specificity than the LDA at both 98% sensitivity (0.37 ± 0.10 versus 0.13 ± 0.13, $p=0.04$) and at 100% sensitivity (0.34 ± 0.08 versus 0.09 ± 0.12, $p=0.03$).

	DF-A	DF-P	ANN	LDA
DF-A		0.48	0.45	0.27
DF-P			0.14	0.12
ANN				0.46
LDA				

of this information flow from the individual features to the decision fuser. The robustness of the algorithm also suggests that decision fusion may be able to reach the asymptotic validation performance with fewer data. This is important for most medical researchers who are starting to collect new databases and for any databases that are expensive to collect. Because our decision-fusion technique needs to estimate only one-dimensional PDFs, which require much fewer data points than multidimensional PDFs, decision fusion needs many fewer data points for training. For this reason, the decision-fusion algorithm may be able to handle typical clinical data sets with missing data, as shown in previous work with decision fusion.[44]

Drawbacks of the decision-fusion algorithm include losing potentially useful feature information by reducing the likelihood-ratio values of the features to a binary value. Although the algorithm loses some feature information in this step, it recovers by optimally fusing the remaining binary feature information from that point forward. In the ideal case, if the true underlying multivariate distribution of the data happens to be known or can be estimated with a high degree of confidence, then the Bayes classifier can take this information into account and is theoretically optimal. However, since the true underlying distribution is almost never known in practice, decision fusion is a good alternative method, especially for small and noisy data sets.

V. CONCLUSIONS

We have developed a decision-fusion classification technique that combines features from heterogeneous data sources. We have demonstrated the technique on both a data set of two different breast imaging modalities and a data set of human-extraced versus computer-extracted findings. With our data, decision fusion always performed as well as or better than the classic classification techniques LDA and ANN. The improvements were all significant for the more challenging Data set C, but not always significant for the less challenging Data set M. Such a statement may not reflect the full diversity of these data sets, which differ in many respects, including linear separability, numbers of cases and features, and feature correlations. Future work will explore the contribution of such factors in order to understand the full potential and limitations of the decision-fusion tech-

nique. In conclusion, the decision-fusion technique showed particular strength in the task of combining groups of weak noisy features for classification.

ACKNOWLEDGMENTS

This work was supported by U.S. Army Breast Cancer Research Program (Grant Nos. W81XWH-05-1-0292 and DAMD17-02-1-0373), and NIH/NCI (Grant Nos. R01 CA95061 and R21 CA93461). We thank Brian Harrawood for the ROC bootstrap code, Anna Bilska-Wolak, Ph.D., and Georgia Tourassi, Ph.D., for insightful discussions, and Andrea Hong, M.D., Jennifer Nicholas, M.D., Priscilla Chyn, and Susan Lim for data collection.

[a]Electronic mail: jonathan.jesneck@duke.edu

[1]J. V. Lacey, Jr., S. S. Devesa, and L. A. Brinton, "Recent trends in breast cancer incidence and mortality," Environ. Mol. Mutagen. **39**, 82-88 (2002).

[2]A. Jemal, T. Murray, E. Ward, A. Samuels, R. C. Tiwari, A. Ghafoor, E. J. Feuer, and M. J. Thun, "Cancer statistics, 2005," Ca-Cancer J. Clin. **55**, 10-30 (2005).

[3]B. Cady and J. S. Michaelson, "The life-sparing potential of mammographic screening," Cancer **91**, 1699-1703 (2001).

[4]J. E. Meyer, D. B. Kopans, P. C. Stomper, and K. K. Lindfors, "Occult breast abnormalities: percutaneous preoperative needle localization," Radiology **150**, 335-337 (1984).

[5]A. L. Rosenberg, G. F. Schwartz, S. A. Feig, and A. S. Patchefsky, "Clinically occult breast lesions: localization and significance," Radiology **162**, 167-170 (1987).

[6]B. C. Yankaskas, M. H. Knelson, J. T. Abernethy, J. T. Cuttino, and R. L. Clark, "Needle localization biopsy of occult lesions of the breast," Radiology **23**, 729-733 (1988).

[7]M. A. Helvie, D. M. Ikeda, and D. D. Adler, "Localization and needle aspiration of breast lesions: complications in 370 cases," AJR, Am. J. Roentgenol. **157**, 711-714 (1991).

[8]L. J. W. Burhenne, S. A. Wood, C. J. D'Orsi, S. A. Feig, D. B. Kopans, K. F. O'Shaughnessy, E. A. Sickles, L. Tabar, C. J. Vyborny, and R. A. Castellino, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," Radiology **215**, 554-562 (2000).

[9]T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center," Radiology **220**, 781-786 (2001).

[10]R. F. Brem, J. Baum, M. Lechner, S. Kaplan, S. Souders, L. G. Naul, and J. Hoffmeister, "Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial," AJR, Am. J. Roentgenol. **181**, 687-693 (2003).

[11]S. V. Destounis, P. DiNitto, W. Logan-Young, E. Bonaccio, M. L. Zuley, and K. M. Willison, "Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience," Radiology **232**, 578-584 (2004).

[12]C. J. Vyborny, "Can computers help radiologists read mammograms?" Radiology **191**, 315-317 (1994).

[13]H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network," Phys. Med. Biol. **42**, 549-567 (1997).

[14]M. A. Gavrielides, J. Y. Lo, and C. E. Floyd, Jr., "Parameter optimization of a computer-aided diagnosis scheme for the segmentation of microcalcification clusters in mammograms," Med. Phys. **29**, 475-483 (2002).

[15]N. Petrick, H. P. Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification," Med. Phys. **23**, 1685-1696 (1996).

[16]N. Petrick, B. Sahiner, H. P. Chan, M. A. Helvie, S. Paquerault, and L. M. Hadjiiski, "Breast cancer detection: Evaluation of a mass-detection algorithm for computer-aided diagnosis — Experience in 263 patients," Radiology **224**, 217-224 (2002).

[17]Y. H. Chang, B. Zheng, and D. Gur, "Computerized identification of suspicious regions for masses in digitized mammograms," Invest. Radiol. **31**, 146-153 (1996).

[18]J. Wei, B. Sahiner, L. M. Hadjiiski, H.-P. Chan, N. Petrick, M. A. Helvie, M. A. Roubidoux, J. Ge, and C. Zhou, "Computer-aided detection of breast masses on full field digital mammograms," Med. Phys. **32**, 2827-2838 (2005).

[19]D. Chen, R. F. Chang, and Y. L. Huang, "Breast cancer diagnosis using self-organizing map for sonography," Ultrasound Med. Biol. **26**, 405-411 (2000).

[20]K. Horsch, M. L. Giger, L. A. Venta, and C. J. Vyborny, "Computerized diagnosis of breast lesions on ultrasound," Med. Phys. **29**, 157-164 (2002).

[21]K. Horsch, M. L. Giger, C. J. Vyborny, and L. A. Venta, "Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography," Acad. Radiol. **11**, 272-280 (2004).

[22]W. Chen, M. L. Giger, L. Lan, and U. Bick, "Computerized interpretation of breast MRI: investigation of enhancement-variance dynamics," Med. Phys. **31**, 1076-1082 (2004).

[23]M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins, "Predicting the clinical status of human breast cancer by using gene expression profiles," Proc. Natl. Acad. Sci. U.S.A. **98**, 11462–11467 (2001).

[24]B. Sahiner, H.-P. Chan, L. M. Hadjiiski, M. A. Roubidoux, C. Paramagul, M. A. Helvie, and C. Zhou, "Multimodality CAD: combination of computerized classification techniques based on mammograms and 3D ultrasound volumes for improved accuracy in breast mass characterization," *Proceedings at Medical Imaging 2004: Image Processing*, San Diego, CA (2004), Vol. 5370, pp. 67-74.

[25]P. Pavlidis, J. Weston, J. Cai, and W. S. Noble, "Learning gene functional classifications from multiple data types," J. Comp. Biol. **9**, 401-411 (2002).

[26]G. R. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble, "A statistical framework for genomic data fusion," Bioinformatics **20**, 2626-2635 (2004).

[27]G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble, "Kernel-based data fusion and its application to protein function prediction in yeast," *Proceedings of the Pacific Symposium on Biocomputing* (2004), World Scientific Press, Kohala Coast, Hawaii, pp. 300-311.

[28]Y. Liao, L. W. Nolte, and L. Collins, "Optimal multisensor decision fusion of mine detection algorithms," Proc. SPIE **5089**, 1252-1260 (2003).

[29]K. Veeramachaneni, L. A. Osadciw, and P. K. Varshney, "An adaptive multimodal biometric management algorithm," IEEE Trans. Syst. Man Cybern. **35**, 344-356 (2005).

[30]M. M. Zheng, S. M. Krishnan, and M. P. Tjoa, "A fusion-based clinical decision support for disease diagnosis from endoscopic images," Comput. Biol. Med. **35**, 259-274 (2005).

[31]H. P. Chan, B. Sahiner, K. L. Lam, N. Petrick, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces," Med. Phys. **25**, 2007-2019 (1998).

[32]M. Kallergi, "Computer-aided diagnosis of mammographic microcalcification clusters," Med. Phys. **31**, 314-326 (2004).

[33]M. K. Markey, J. Y. Lo, and C. E. Floyd, Jr., "Differences between computer-aided diagnosis of breast masses and that of calcifications," Radiology **223**, 489-493 (2002).

[34]M. Heath, K. W. Bowyer, and D. Kopans, "Current status of the digital database for screening mammography," in *Digital Mammography*, edited by N. Karssemeijer, M. Thijssen, and J. Hendriks (Kluwer Academic, Dordrecht, 1998), pp. 457-460.

[35]BI-RADS, American College of Radiology Breast Imaging - Reporting and Data System 3rd ed., American College of Radiology, Reston, VA, 1998.

[36]A. S. Hong, E. L. Rosen, M. S. Soo, and J. A. Baker, "BI-RADS for sonography: Positive and negative predictive values of sonographic features," AJR, Am. J. Roentgenol. **184**, 1260-1265 (2005).

[37]J. L. Jesneck, J. Y. Lo, and J. A. Baker, "A computer aid for diagnosis of breast mass lesions using both mammographic and sonographic BI-RADS descriptors," Radiology (in press).

[38]R. R. Tenney and N. R. Sandell, Jr., *Detection with Distributed Sensors*, Proceedings of the IEEE Conference on Decision and Control, **1**, 501-510 (1980).

[39]Z. Chair and P. K. Varshney, "Optimal data fusion in multiple sensor detection systems," IEEE Trans. Aerosp. Electron. Syst. **AES-22**, 98 (1986).

[40]A. R. Reibman and L. W. Nolte, "Optimal detection and performance of distributed sensor systems," IEEE Trans. Aerosp. Electron. Syst. AES-**23**, 24-30 (1987).
[41]B. V. Dasarathy, "Decision fusion strategies in multisensor environments," IEEE Trans. Syst. Man Cybern. **21**, 1140-1154 (1991).
[42]Y. Liao, "Distributed decision fusion in signal detection-A robust approach," Ph. D. thesis, Duke University, 2005.
[43]R. Niu, P. K. Varshney, M. Moore, and D. Klamer, "Decision fusion in a wireless sensor network with a large number of sensors," *Proceedings of the Seventh International Conference on Information Fusion, FUSION 2004*, Stockholm, Sweden (2004), Vol. 1, International Society of Information Fusion, p. 21.
[44]A. O. Bilska-Wolak and C. E. Floyd, Jr., "Tolerance to missing data using a likelihood ratio based classifier for computer-aided classification of breast cancer," Phys. Med. Biol. **49**, 4219-4237 (2004).
[45]A. O. Bilska-Wolak, C. E. Floyd, Jr., L. W. Nolte, and J. Y. Lo, "Application of likelihood ratio to classification of mammographic masses; performance comparison to case-based reasoning," Med. Phys. **30**, 949-958 (2003).
[46]A. O. Bilska-Wolak, C. E. Floyd, Jr., J. Y. Lo, and J. A. Baker, "Computer aid for decision to biopsy breast masses on mammography: validation on new cases," Acad. Radiol. **12**, 671-680 (2005).
[47]H. L. VanTrees, *Detection, Estimation, and Modulation Theory (Part I)* (Wiley, New York, 1968).
[48]E. Drakopoulos and C.-C. Lee, "Optimum multisensor fusion of correlated local decisions," IEEE Trans. Aerosp. Electron. Syst. **27**, 593-606 (1991).
[49]Y. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," Radiology **201**, 745-750 (1996).
[50]B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap* (Chapman and Hall, New York, 1993).

Published online before print June 11, 2007
(*Radiology* 2007, 10.1148/radiol.2442060712)

Breast Imaging

Breast Mass Lesions: Computer-aided Diagnosis Models with Mammographic and Sonographic Descriptors[1]

Jonathan L. Jesneck, PhD, Joseph Y. Lo, PhD, and Jay A. Baker, MD

[1] From the Department of Biomedical Engineering (J.L.J., J.Y.L.) and Duke Advanced Imaging Labs, Department of Radiology (J.L.J., J.Y.L., J.A.B.), Duke University Medical Center, 2424 Erwin Rd, Suite 302, Durham, NC 27705. Received April 23, 2006; revision requested June 23; revision received July 24; accepted August 29; final version accepted November 15. Supported by U.S. Army Breast Cancer Research Program W81XWH-05-1-0292 and DAMD17-02-1-0373, and NIH/NCI R01 CA95061 and R21 CA93461. **Address correspondence to** J.L.J. (e-mail: *jonathan.jesneck@duke.edu*).

ABSTRACT

Purpose: To retrospectively develop and evaluate computer-aided diagnosis (CAD) models that include both mammographic and sonographic descriptors.

Materials and Methods: Institutional review board approval was obtained for this HIPAA-compliant study. A waiver of informed consent was obtained. Mammographic and sonographic examinations were performed in 737 patients (age range, 17–87 years), which yielded 803 breast mass lesions (296 malignant, 507 benign). Radiologist-interpreted features from mammograms and sonograms were used as input features for linear discriminant analysis (LDA) and artificial neural network (ANN) models to differentiate benign from malignant lesions. An LDA with all the features was compared with an LDA with only stepwise-selected features. Classification performances were quantified by using receiver operating characteristic (ROC) analysis and were evaluated in a train, validate, and retest scheme. On the retest set, both LDAs were compared with radiologist assessment score of malignancy.

Results: Both the LDA and ANN achieved high classification performance with cross validation (area under the ROC curve [A_z] = 0.92 ± 0.01 [standard deviation] and $_{0.90}A_z$ = 0.54 ± 0.08 for

LDA, A_z = 0.92 ± 0.01 and $_{0.90}A_z$ = 0.55 ± 0.08 for ANN). Results of both models generalized well to the retest set, with no significant performance differences between the validate and retest sets (P > .1). On the retest set, there were no significant performance differences between LDA with all features and LDA with only the stepwise-selected features (P > .3) and between either LDA and radiologist assessment score (P > .2).

Conclusion: Results showed that combining mammographic and sonographic descriptors in a CAD model can result in high classification and generalization performance. On the retest set, LDA performance matched radiologist classification performance.

INTRODUCTION

TOP
ABSTRACT
INTRODUCTION
MATERIALS AND METHODS
RESULTS
DISCUSSION
ADVANCES IN KNOWLEDGE
References

Because of low specificity at mammography, many women undergo unnecessary breast biopsy. As many as 65%–85% of breast biopsies are performed in benign lesions (1–3). Unnecessary biopsy not only increases the cost of mammographic screening (4) but also subjects patients to avoidable emotional and physical burdens.

To improve the accuracy of mammography, computer aids have become available to help radiologists detect (5–8) and diagnose (9–12) suspicious breast lesions. Some study results (13,14) have shown that use of such computer-aided diagnosis (CAD) systems has increased overall diagnostic sensitivity and specificity. Lesions determined to be very likely benign may be recommended for short-term follow-up rather than biopsy (13,14).

CAD models often involve breast morphologic descriptors of the Breast Imaging Reporting and Data System (BI-RADS) lexicon. BI-RADS was developed by the American College of Radiology to standardize the interpretation of mammograms (15–17). Originally, BI-RADS was applied to only mammography, but the crucial adjunct role of sonography has recently led the American College of Radiology to develop a BI-RADS lexicon for breast sonography as well (18). Sonographic BI-RADS is a useful tool to help standardize the characterization of sonographic lesions (18,19) and facilitate clinician communication.

Until recently, the primary clinical role for sonography has been to aid in distinguishing simple cysts from solid masses, as well as to direct aspirations, wire localizations, and biopsies. Several authors (20–24) have investigated the role of sonography in helping to differentiate malignant from benign breast lesions. There also have been many CAD studies (25–33) of breast sonography, which are based on image features automatically extracted by using computer vision algorithms. To the best of our knowledge, there has not yet been a published study with either the standardized BI-RADS sonographic findings as the basis of a predictive model or the combination of BI-RADS mammographic and sonographic findings for that purpose. Thus, the purpose of our study was to retrospectively develop and evaluate CAD models that involve both mammographic and sonographic descriptors.

MATERIALS AND METHODS

TOP
ABSTRACT
INTRODUCTION

Lesions and Patients

Institutional review board approval was obtained for this Health Insurance Portability and Accountability Act–compliant study. A

▪ MATERIALS AND METHODS
RESULTS
DISCUSSION
ADVANCES IN KNOWLEDGE
References

waiver of informed consent was obtained. The lesions used in this study were an extension of an original 403-lesion data set described in detail in a previous study (34). They were collected between 2000 and 2005 at our institution. The data set included 803 lesions, of which 296 were malignant and 507 were benign, and 389 were palpable and 414 were nonpalpable. There were 737 patients whose ages ranged from 17 to 87 years, with a median age of 50 years. The same inclusion and exclusion criteria as described previously (34) applied to this data set. Lesions were selected from those recommended for biopsy and were included in the study if the lesions corresponded to solid masses on sonograms and if both mammographic and sonographic images taken before the biopsy were available for review. Any complicated cysts were excluded from consideration. All cases were re-reviewed by one of four breast radiologists (including J.A.B.) who were blinded to the original report.

Features Used

All patients underwent both mammography and sonography. The mammographic examination consisted of both craniocaudal and mediolateral oblique views, with additional true lateral and spot compression magnification when clinically appropriate. Sonographic images were acquired in both radial and antiradial projections, with and without caliper measurements. Additional gray-scale images were obtained in almost all patients to better depict the lesion. Doppler, color Doppler, and power Doppler images were not part of the routine imaging protocol but were reviewed when available. One of four dedicated breast radiologists (including J.A.B.) with 6–11 years of experience used BI-RADS lexicon to describe the lesions, as described previously (34). Information about patient physical examination findings, family history of breast cancer, and personal history of breast malignancy was available to each radiologist to reproduce a realistic clinical situation. The radiologist was blinded to the histologic diagnosis during the evaluation.

Of the total 39 features, 13 were mammographic BI-RADS features, 13 were sonographic BI-RADS features, six were sonographic features suggested by Stavros et al (20), four were other sonographic features, and three were patient history features. The 13 mammographic BI-RADS features were mass size, parenchyma density, mass margin, mass shape, mass density, calcification number of particles, calcification distribution, calcification description, architectural distortion, associated findings, special cases (as defined by the BI-RADS lexicon: asymmetric tubular structure, intramammary lymph node, global asymmetry, and focal asymmetry), comparison with findings at prior examination, and change in mass size. The 13 sonographic BI-RADS features were radial diameter, antiradial diameter, anteroposterior diameter, background tissue echo texture, mass shape, mass orientation, mass margin, lesion boundary, echo pattern, posterior acoustic features, calcifications within mass, special cases (as defined by the BI-RADS lexicon: clustered microcysts, complicated cysts, mass in or on skin, foreign body, intramammary lymph node, and axillary lymph node), and vascularity. The six features suggested by Stavros et al (20) were mass shape, mass margin, acoustic transmission, thin echo pseudocapsule, mass echogenicity, and calcifications. The four other sonographic mass descriptors were edge shadow, cystic component, and two mammographic BI-RADS descriptors applied to sonography—mass shape (oval and lobulated are separate descriptors) and mass margin (replaces sonographic descriptor *angular* with *obscured*). The three patient history features were family history, patient age, and indication for sonography.

In addition to the BI-RADS and Stavros et al descriptors, the radiologists also recorded their assessment about the malignancy of the lesion as an integer ranging from 0 for unquestionably benign to 100 for unquestionably malignant. This assessment rating was not used as an input to the CAD models but rather as a comparison to the models' output for classification performance.

Predictive Modeling, Sampling, and Feature Selection

For models in this study, we (J.L.J. and J.Y.L. by consensus) used linear discriminant analysis (LDA) and artificial neural networks (ANNs). The LDA was a Fisher linear discriminant. The ANNs were three-layer (one hidden layer), feed-forward, and error back-propagation models. These are the most common methods used in many previous studies by our group, as well as the rest of the field.

To assess the usefulness and risk of using CAD models in the clinic, it is crucial to have a good estimate of their performance in future cases (or generalization). For limited data and more complicated models, the traditional method of cross validation could still pose a danger of optimistically biasing the testing performance; it is common to optimize certain global parameters (such as feature selection for the LDA or number of hidden nodes of an ANN) to maximize cross-validation performance. With cross validation, one is able to use knowledge of all the data to make modeling decisions, whereas with generalization such information is not available for yet unseen future cases. Therefore, optimizing the models for cross-validation performance could lead to reduced generalization performance.

To avoid these overfitting pitfalls and to better estimate generalization ability of each model, we used a train, validate, and retest scheme. In this scheme, the data set was divided into sets: a train and validate set and a retest set. The retest set was not used until the models were finalized, so as not to influence any of the modeling process. All modeling decisions were made only on the train and validate set. The model parameters were optimized to maximize cross validation on the train and validate set. Once the model's parameter values were set, the model was then trained on the entire train and validate set. The trained model was then applied to the retest set.

In particular, for our data set of 803 lesions, we chose the first 500 lesions in chronologic order for the train and validate set and the remaining 303 lesions for the retest set. We chose architecture and parameter settings for the ANN to optimize its cross-validation performance on the train and validate set. Once the modeling decisions had been made, we trained the LDA and ANN on all the lesions in the train and validate set to determine a single, final set of weights, which were then applied to the retest set.

In addition to aiding model training and assessment, the train, validate, and retest scheme can also reduce bias in feature selection. Using this scheme, we investigated the effect of feature selection on the generalization performance of an LDA. Using only the validate set, we performed stepwise feature selection. We then used these selected features to train an LDA on the train and validate set. We then applied the trained LDA model to the retest set. Finally, on the retest set, we compared the generalization performance of the LDA with only the stepwise-selected features and that of the LDA with all the features.

Classifier Performance Evaluation and Statistical Analysis

To use the LDA or ANN model as a diagnostic aid, one could select a threshold value, so that lesions with output values less than the threshold would be considered very likely benign and therefore candidates for follow-up rather than biopsy. Those lesions with model outputs greater than the threshold would be considered suspicious for malignancy and recommended for biopsy.

Varying the threshold value results in a trade-off between sensitivity and specificity. The entire range of sensitivity and specificity values for a classifier is illustrated by using the receiver operating characteristic (ROC) curve (35,36). To quantify a classifier's performance, we (J.L.J. and J.Y.L. by consensus) used the following five summary measures of the ROC curve: area under the ROC curve (A_z), the partial area ($_{0.90}A_z$), and the specificity, positive predictive value, and negative predictive value for a given sensitivity level. A_z represents the average specificity across all sensitivities and ranges from 0.5 (chance performance) to 1.0 (perfect performance). Because

high sensitivity is essential for a classification task, a more relevant performance measure is $_{0.90}A_z$, which represents the average specificity performance of the classifier at sensitivities from 90% to 100%.

Whereas the two previous measures provide an overall summary of performance, the remaining three are clinically relevant measures that correspond to a single threshold value, which for breast cancer applications is usually chosen to deliver nearly perfect sensitivity, such as 98% (37,38). Note that for this data set, the actual positive predictive value of the clinical decision to refer to biopsy was 37%, which is typical of our institution. Because our study included only biopsy-verified lesions, sensitivity was 100% and specificity was 0% for cancer detection by definition.

These classifier performance metrics allowed us to compare classifier performance statistically. We used the nonparametric bootstrap method (39) to measure the means and variances of the classification metric values, as well as to compare metric results of the two models for statistical significance. Although we assumed statistical independence of the lesions for modeling, 8% (66 of 803) of the BI-RADS data set included multiple lesions per patient. To adjust for clustering of data values, we used cross validation by patient, which ensured that no lesions from the same patient appeared in more than one of the train, validate, and retest sets. A *P* value of less than .05 was considered to indicate a significant difference.

RESULTS

Generalization between Validating and Retesting

The LDA achieved high classification performance, with A_z = 0.92 ± 0.01 and $_{0.90}A_z$ = 0.54 ± 0.08 on the train and validate set and A_z = 0.92 ± 0.02 and $_{0.90}A_z$ = 0.52 ± 0.08 on the retest set (Table 1). Results of the LDA generalized well; there were no significant differences between the performance metric results of the validate set and those of the retest set (P > .10). In addition to the entire ROC curves of the LDA performance, results with individual thresholds also generalized well. The same threshold value determined similar true-positive fraction (sensitivity) and false-positive fraction (1 – specificity) operating points in the high-sensitivity region on both ROC curves (Table 2).

View this table: [in this window] [in a new window]	**Table 1. Classification Performance of LDA as Measured with ROC Curve**

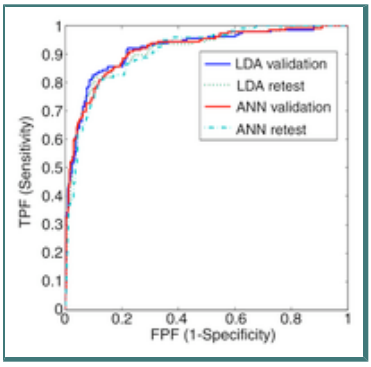
View this table: [in this window] [in a new window]	**Table 2. Generalization of LDA Thresholds**

The ANN also performed well, achieving A_z = 0.92 ± 0.01 and $_{0.90}A_z$ = 0.55 ± 0.08 on the validate set and A_z = 0.91 ± 0.02 and $_{0.90}A_z$ = 0.57 ± 0.06 on the retest set. The ANN performed

comparably on the validate and retest sets, with no significant differences in either metric (P > .10).

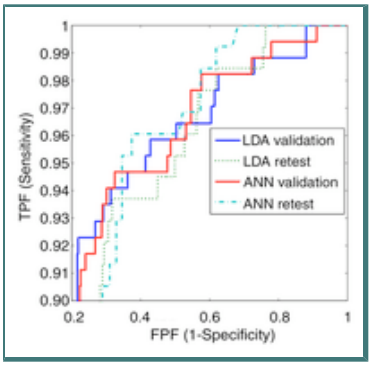
Comparison of LDA and ANN Performance

The two types of models, LDA and ANN, had similar performances on both the validate and retest sets; the differences were not significant (P > .10). In the interest of brevity, tables with the results of ANN performance are not included in our study because of their close similarity to tables with LDA performance results. ROC curves for the LDA and ANN in both testing paradigms (Fig 1) showed that discrepancies among the curves were minor, and the curves overlap each other with essentially indistinguishable classification performance.



View larger version (27K):
[in this window]
[in a new window]
[Download PPT slide]

Figure 1a: (a) Full ROC curves for classifier performance: validate set versus retest set. **(b)** Partial ROC curves for classifier performance: cross validation versus retest set. Results of LDA and ANN generalized well on retest data set, as shown by their overlapping ROC curves. Validation ROC curves (solid curves) lie close to retest ROC curves (dashed curves). LDA and ANN had virtually indistinguishable classification performances. *FPF* = false-positive fraction, *TPF* = true-positive fraction.



View larger version (30K):
[in this window]
[in a new window]
[Download PPT slide]

Figure 1b: (a) Full ROC curves for classifier performance: validate set versus retest set. **(b)** Partial ROC curves for classifier performance: cross validation versus retest set. Results of LDA and ANN generalized well on retest data set, as shown by their overlapping ROC curves. Validation ROC curves (solid curves) lie close to retest ROC curves (dashed curves). LDA and ANN had virtually indistinguishable classification performances. *FPF* = false-positive fraction, *TPF* = true-positive fraction.

Feature Selection and Generalization of Simplified Model

Performance of stepwise feature selection for the LDA resulted in the following 14 features: patient age, calcification distribution, calcification description, associated findings, comparison with findings at prior examination, anteroposterior diameter, indication for sonography, Stavros et al mass shape, mammographic BI-RADS mass margin, edge shadow, cystic component,

sonographic lesion boundary, surrounding tissue effects, and sonographic special findings. An LDA with only these stepwise-selected features performed comparably to the LDA with all the features, with no significant difference (A_z = 0.92 ± 0.02 vs 0.91 ± 0.02, respectively; P > .3). A table with performance results of the LDA with stepwise-selected features was not included in our study because of its close similarity to the table with results of the LDA with all features.

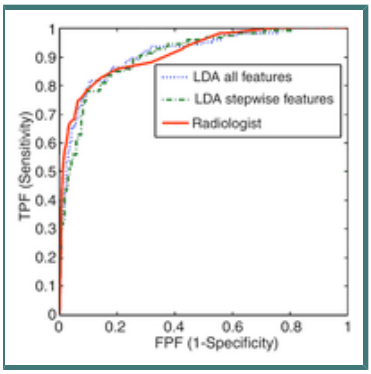
Comparing LDA to Radiologist Assessment of Malignancy

Like the LDA, radiologist assessment also achieved high classification performance on the retest set (Table 3), with A_z = 0.92 ± 0.02 and $_{0.90}A_z$ = 0.52 ± 0.06 on the retest set. There were no significant differences between any of the performance metric results of the LDA and radiologist assessment (P > .2). For example, on this retest data set, the LDA and radiologists performed with similar negative predictive values (97% ± 1 vs 98% ± 1, respectively; P = .25).

View this table:
[in this window]
[in a new window]

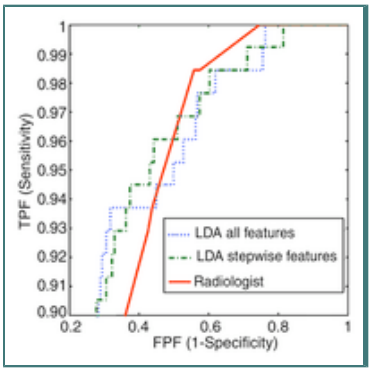
Table 3. LDA versus Radiologist Assessment on Retest Set

With regard to ROC curves for the LDA with all features, the LDA with the stepwise-selected features, and radiologist assessment of malignancy (Fig 2), there were no significant differences in any of the performance metric results among the three ROC curves (P > .2). Although the radiologist curve crossed the LDA curves several times, even at the points of greater divergence, the differences were not significant (P > .2). In a lesion in which the LDA and radiologist disagreed (Fig 3), the LDA correctly classified the lesion as benign.



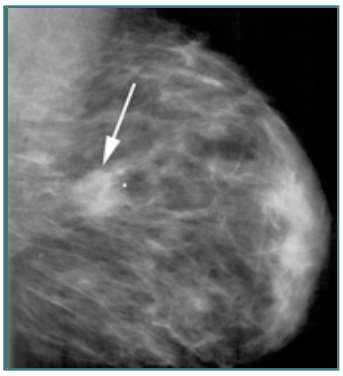
View larger version (27K):
[in this window]
[in a new window]
[Download PPT slide]

Figure 2a: (a) Full ROC curves: LDA versus radiologist on retest set. **(b)** Partial ROC curves: LDA versus radiologist on retest set. ROC curves for LDA with all features, for LDA with stepwise-selected features, and for radiologist assessment of malignancy. In retesting, LDA, both with all features and with only stepwise-selected features, performed similarly to radiologists. There were no significant differences in any performance metric results among the three ROC curves (P > .2). Although the radiologist curve crossed LDA curves several times, even at points of greater divergence, differences were not significant (P > .2). *FPF* = false-positive fraction, *TPF* = true-positive fraction.



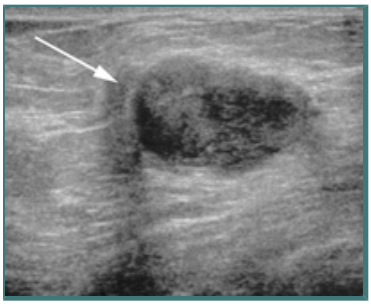
View larger version (32K):
[in this window]
[in a new window]
[Download PPT slide]

Figure 2b: (a) Full ROC curves: LDA versus radiologist on retest set. **(b)** Partial ROC curves: LDA versus radiologist on retest set. ROC curves for LDA with all features, for LDA with stepwise-selected features, and for radiologist assessment of malignancy. In retesting, LDA, both with all features and with only stepwise-selected features, performed similarly to radiologists. There were no significant differences in any performance metric results among the three ROC curves ($P > .2$). Although the radiologist curve crossed LDA curves several times, even at points of greater divergence, differences were not significant ($P > .2$). *FPF* = false-positive fraction, *TPF* = true-positive fraction.



View larger version (137K):
[in this window]
[in a new window]
[Download PPT slide]

Figure 3a: (a) Mediolateral oblique mammogram in 26-year-old patient demonstrates ill-defined, oval-shaped, equal-density mass (arrow) in posterior left breast. Radiopaque marker immediately anterior to mass indicates that this mass was palpable. **(b)** Sonogram in same patient demonstrates oval, circumscribed mass (arrow) with parallel orientation and no posterior acoustic features. Histopathologic diagnosis indicated that this lesion was necrotic breast tissue. Follow-up examination findings confirmed no interval change 2 years after biopsy. LDA considered this lesion relatively benign, with a score of 0.33 of 1.00, whereas radiologist considered it more indicative of malignancy, with a score of 85 of 100.

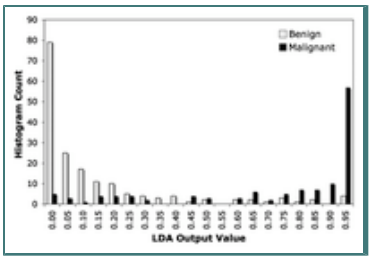


View larger version (141K):
[in this window]
[in a new window]
[Download PPT slide]

Figure 3b: (a) Mediolateral oblique mammogram in 26-year-old patient demonstrates ill-defined, oval-shaped, equal-density mass (arrow) in posterior left breast. Radiopaque marker immediately anterior to mass indicates that this mass was palpable. **(b)** Sonogram in same patient demonstrates oval, circumscribed mass (arrow) with parallel orientation and no posterior acoustic features. Histopathologic diagnosis indicated that this lesion was necrotic breast tissue. Follow-up examination findings confirmed no interval change 2 years after biopsy. LDA considered this lesion relatively benign, with a score of 0.33 of 1.00, whereas radiologist considered it more indicative of malignancy, with a score of 85 of 100.

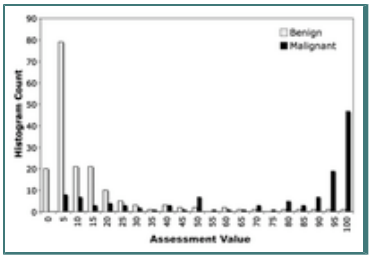
Histograms of the LDA output and radiologist assessment values for the retest set (Fig 4) showed

that the values for the benign lesions (such as in Fig 5) tended to be on the left side of the histogram plot with values around zero. Values for the malignant lesions (such as in Fig 6) were concentrated on the right side of the plots, around 1 for the LDA values and around 100 for radiologist assessment values. There were few values in the center regions compared with those on the extremes.



View larger version (21K):
[in this window]
[in a new window]
[Download PPT slide]

Figure 4a: Histograms of **(a)** LDA output values and **(b)** radiologist assessment. Histogram counts for truly benign lesions are shown in gray, and those for truly malignant lesions are shown in black. For classification, a threshold would be applied to LDA output, so that output values below the threshold would be designated benign and those above it would be designated malignant.



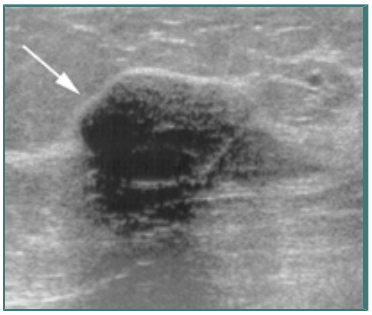
View larger version (20K):
[in this window]
[in a new window]
[Download PPT slide]

Figure 4b: Histograms of **(a)** LDA output values and **(b)** radiologist assessment. Histogram counts for truly benign lesions are shown in gray, and those for truly malignant lesions are shown in black. For classification, a threshold would be applied to LDA output, so that output values below the threshold would be designated benign and those above it would be designated malignant.

View larger version (87K):
[in this window]
[in a new window]

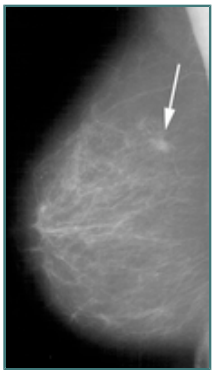
Figure 5a: (a) Mediolateral oblique mammogram in 52-year-old patient demonstrates oval, well-circumscribed, equal-density mass (arrow) in superior left breast. **(b)** Sonogram in same patient demonstrates oval, hypoechoic solid mass (arrow) with circumscribed margins, parallel orientation, and posterior acoustic shadowing. Histopathologic results indicated benign fibroadenoma. Both LDA and radiologist correctly considered this lesion very benign, with scores of 0.02 of 1.00 and 0 of 100, respectively.

[Download PPT slide]



View larger version (142K):
[in this window]
[in a new window]
[Download PPT slide]

Figure 5b: (a) Mediolateral oblique mammogram in 52-year-old patient demonstrates oval, well-circumscribed, equal-density mass (arrow) in superior left breast. **(b)** Sonogram in same patient demonstrates oval, hypoechoic solid mass (arrow) with circumscribed margins, parallel orientation, and posterior acoustic shadowing. Histopathologic results indicated benign fibroadenoma. Both LDA and radiologist correctly considered this lesion very benign, with scores of 0.02 of 1.00 and 0 of 100, respectively.



View larger version (74K):
[in this window]
[in a new window]
[Download PPT slide]

Figure 6a: (a) Mediolateral oblique mammogram in 57-year-old patient demonstrates ill-defined, irregularly shaped, equal-density mass (arrow) in superior right breast. **(b)** Sonogram in same patient demonstrates ill-defined, irregularly shaped mass (arrow) with posterior acoustic shadowing and without parallel orientation. Histopathologic diagnosis indicated that this malignant lesion was invasive ductal carcinoma. Both LDA and radiologist correctly considered this lesion very malignant, with scores of 0.99 of 1.00 and 95 of 100, respectively.

View larger version (170K):

Figure 6b: (a) Mediolateral oblique mammogram in 57-year-old patient demonstrates ill-defined, irregularly shaped, equal-density mass (arrow) in superior right breast. **(b)** Sonogram in same patient demonstrates ill-defined, irregularly shaped mass (arrow) with posterior acoustic shadowing and without parallel orientation. Histopathologic diagnosis indicated that this malignant lesion was invasive ductal carcinoma. Both LDA and radiologist correctly considered this lesion very malignant, with scores of 0.99 of 1.00 and 95 of 100, respectively.

[in this window]
[in a new window]
[Download PPT slide]

ROC curves for generalization performance (Fig 2) suggest that radiologists may be able to achieve considerable improvements in performance by shifting their diagnostic performance to a more desirable operating point on the ROC curve. For example, they may perform at 52% specificity, 60% positive predictive value, and 98% negative predictive value by adjusting their mental threshold to reduce their sensitivity slightly to 98% sensitivity, which would have resulted in the delayed diagnosis of 2% of cancers that may be identified by using interval change at a short-term follow-up diagnostic study. Likewise, if the radiologists were hypothetically to adopt all the recommendations of the computer model, they could have perhaps attained 37% specificity, 53% positive predictive value, and 97% negative predictive value at that same 98% sensitivity level.

DISCUSSION

TOP
ABSTRACT
INTRODUCTION
MATERIALS AND METHODS
RESULTS
DISCUSSION
ADVANCES IN KNOWLEDGE
References

To the best of our knowledge, our study is the first CAD study not only to use sonographic BI-RADS features but also to combine BI-RADS features of sonography with those of mammography. In addition, to justify the clinical use of a CAD system on new patients, it is important to estimate its generalization performance. We have estimated the generalization performance of both LDA and ANN models on our data set by using a train, validate, and retest scheme on our data set. There was good evidence of generalization for the LDA and ANN because there was no decrease in performance from the validation curves to the retest curves.

The LDA and ANN had virtually indistinguishable classification performance, which indicated that the BI-RADS data were highly linear. In general, such results would support the use of the LDA model, which is simpler than the nonlinear ANN and therefore less likely to be susceptible to overtraining problems. Our study results, however, demonstrated that there were no problems with overtraining, as both models performed similarly during the retesting phase.

Because CAD systems typically give as output a range of values, applying a certain threshold to the output determines the operating point (sensitivity and specificity settings) at which the clinical decision is made. Knowing the CAD operating point helps the clinician incorporate it into an overall diagnostic decision. We have shown that results with LDA thresholds from the validation ROC curve generalized well to the retest ROC curve in the clinically important high-sensitivity region, which suggests that these threshold values could be used clinically with the LDA on future lesions.

Because the task of collecting many features can be cumbersome, we investigated CAD performance with only a subset of the features by performing stepwise feature selection. Of the 14 selected features, three also had been found to have high malignancy predictive value from a previous study (34): Stavros et al mass shape, mammographic mass margin, and sonographic lesion boundary. To ensure that the selected features were adequate to allow the CAD system to achieve good generalization on new lesions, a train, validate, and retest scheme was required. Only the train and validate set was used to select the features, which were then tested in a CAD model on the retest set. LDA with only the 14 stepwise-selected features performed just as well as an LDA with all 37 features. The small number of features required for good performance suggests that this CAD model may be able to offer the benefit of having a second reader without

greatly slowing workflow. Similar feature definitions caused some features to be collinear. While data collinearity does potentially bias the selected model to be optimistic, the rigorous use of cross validation followed by completely independent retesting demonstrated that the results did generalize without optimistic bias.

LDA results distinguished benign from malignant lesions no differently than did radiologist assessment scores for our data set. The generalization performance results suggest that radiologists may be able to achieve a more desirable operating point on their ROC curve by adjusting their mental threshold to have slightly lower sensitivity but much higher specificity. If the radiologists were to adopt all the recommendations of the computer model, they could substantially increase specificity while maintaining a high sensitivity level.

The radiologists in this study were experienced dedicated breast imagers. It is hoped that less-specialized radiologists using such a system could improve their diagnostic performance closer to that of breast specialists. In practice, it remains to be determined how radiologists would use the results from such computer models, in particular whether they would modify a recommendation for biopsy to a recommendation for short-term follow-up in those lesions deemed to be very likely benign. It also remains unknown whether the 2% of cancers mistakenly referred to follow-up would prove to remain early stage, such as with the current clinical practice of following up probably benign lesions.

There were limitations to our study. The BI-RADS data collection included multiple lesions per patient for 66 of 803 lesions. Our study criteria included solid masses rather than cysts and the use of only biopsy-proved lesions. Additionally, radiologists allowed mammograms to influence their recording of the sonographic features, because they analyzed mammograms immediately before sonograms. The study was organized in this manner to better reflect actual clinical practice in which the mammogram is obtained immediately prior to the sonogram and decisions are made by using all available data. They also could have shifted their diagnostic sensitivity and specificity levels from their usual clinical levels because they were aware that the lesion diagnoses had been resolved, and therefore, their assessment ratings did not directly affect patient care.

In conclusion, the results of model classification and generalization performance on our data set suggest that the models could be used as a CAD system for future mass lesions. Because the results with LDA threshold values generalized well, the desired operating point on the ROC curve could be set for future lesions, which increases the usefulness of the CAD system. Because the stepwise-selected features were adequate for good classification and generalization, they could be used in a CAD system that would require only minimal feature collection. In our study, we were not trying to improve diagnostic accuracy of dedicated breast imagers but rather to offer a tool to radiologists to allow a substantial decrease in the number of unnecessary benign breast biopsies while minimizing the number of delayed breast cancer diagnoses.

ADVANCES IN KNOWLEDGE

- Both mammographic and sonographic Breast Imaging Reporting and Data System descriptors are useful in a computer-aided diagnosis (CAD) system for differentiating malignant from benign breast masses with high performance (area under the receiver operating characteristic curve, 0.92 ± 0.02).
- Results with this multimodal CAD system generalized well to new lesions, an important step for the consideration of incorporating a CAD system into clinical use.

ACKNOWLEDGMENTS

We thank David DeLong, PhD, for help with statistical analysis, Brian Harrawood, BA, for the ROC bootstrap code, Carey Floyd, Jr, PhD (deceased), and Georgia Tourassi, PhD, for insightful discussions, and Andrea Hong, MD, and Priscilla Chyn, MD, for data collection.

FOOTNOTES

Abbreviations: ANN = artificial neural network • A_z = area under the ROC curve • BI-RADS = Breast Imaging Reporting and Data System • CAD = computer-aided diagnosis • LDA = linear discriminant analysis • ROC = receiver operating characteristic

Authors stated no financial relationship to disclose.

Author contributions: Guarantors of integrity of entire study, J.L.J., J.Y.L.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; manuscript final version approval, all authors; literature research, J.L.J.; clinical studies, J.A.B.; statistical analysis, J.L.J., J.Y.L.; and manuscript editing, all authors

References

1. Kopans DB. The positive predictive value of mammography. AJR Am J Roentgenol 1992;158:521–526.[Free Full Text]
2. Ciatto S, Cataliotti L, Distante V. Nonpalpable lesions detected with mammography: review of 512 consecutive cases. Radiology 1987;165:99–102.[Abstract]
3. Meyer JE, Eberlein TJ, Stomper PC, Sonnenfeld MR. Biopsy of occult breast lesions: analysis of 1261 abnormalities. JAMA 1990;263:2341–2343.[Abstract]
4. Cyrlak D. Induced costs of low-cost screening mammography. Radiology 1988;168:661–663.[Abstract]
5. Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000;215:554–562.[Abstract/Free Full Text]
6. Zheng B, Chang YH, Wang XH, Good WF, Gur D. Feature selection for computerized mass detection in digitized mammograms by using a genetic algorithm. Acad Radiol 1999;6:327–332.[Medline]
7. Qian W, Clarke LP, Song D, Clark RA. Digital mammography: hybrid four-channel wavelet transform for microcalcification segmentation. Acad Radiol 1998;5:354–364.[Medline]
8. Qian W, Li L, Clarke L, Clark RA, Thomas J. Digital mammography: comparison of adaptive and nonadaptive CAD methods for mass detection. Acad Radiol 1999;6:471–480.[Medline]
9. Chan HP, Sahiner B, Helvie MA, et al. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. Radiology 1999;212:817–827.[Abstract/Free Full Text]
10. Chan HP, Sahiner B, Lam KL, et al. Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces. Med Phys 1998;25:2007–2019.[Medline]
11. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol 1999;6:22–33.[Medline]
12. Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K. Automated computerized

classification of malignant and benign masses on digitized mammograms. Acad Radiol 1998;5:155–168.[Medline]
13. Baker JA, Kornguth PJ, Lo JY, Williford ME, Floyd CE Jr. Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. Radiology 1995;196:817–822.[Abstract]
14. Lo JY, Baker JA, Kornguth PJ, Floyd CE Jr. Effect of patient history data on the prediction of breast cancer from mammographic findings with artificial neural networks. Acad Radiol 1999;6:10–15.[Medline]
15. Kopans DB. Standardized mammography reporting. Radiol Clin North Am 1992;30:257–264.[Medline]
16. D'Orsi CJ, Kopans DB. Mammographic feature analysis. Semin Roentgenol 1993;28:204–230.[Medline]
17. American College of Radiology. Breast Imaging-Reporting and Data System (BI-RADS). 3rd ed. Reston, Va: American College of Radiology, 1998.
18. American College of Radiology. Ultrasound. In: Breast Imaging-Reporting and Data System atlas (BI-RADS atlas). 4th ed. Reston, Va: American College of Radiology, 2003.
19. Mendelson EB, Berg WA, Merritt CR. Toward a standardized breast ultrasound lexicon, BI-RADS: ultrasound. Semin Roentgenol 2001;36:217–225.[Medline]
20. Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker S, Sisney G. Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. Radiology 1995;196:123–134.[Abstract]
21. Rahbar G, Sie AC, Hansen GC, et al. Benign versus malignant solid breast masses: US differentiation. Radiology 1999;213:889–894.[Abstract/Free Full Text]
22. Jackson VP. The role of US in breast imaging. Radiology 1990;177:305–311.[Medline]
23. Jackson VP. Management of solid breast nodules: what is the role of sonography? Radiology 1995;196:14–15.[Medline]
24. Zonderland HM, Coerkamp EG, Hermans J, van de Vijver MJ, van Voorthuisen AE. Diagnosis of breast cancer: contribution of US as an adjunct to mammography. Radiology 1999;213:413–422.[Abstract/Free Full Text]
25. Chang RF, Kuo WJ, Chen DR, Huang YL, Lee JH, Chou YH. Computer-aided diagnosis for surgical office-based breast ultrasound. Arch Surg 2000;135:696–699.[Abstract/Free Full Text]
26. Chen D, Chang RF, Huang YL. Breast cancer diagnosis using self-organizing map for sonography. Ultrasound Med Biol 2000;26:405–411.[Medline]
27. Giger ML. Computerized analysis of images in the detection and diagnosis of breast cancer. Semin Ultrasound CT MR 2004;25:411–418.[Medline]
28. Horsch K, Giger ML, Vyborny CJ, Venta LA. Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography. Acad Radiol 2004;11:272–280.[Medline]
29. Drukker K, Giger ML, Vyborny CJ, Mendelson EB. Computerized detection and classification of cancer on breast ultrasound. Acad Radiol 2004;11:526–535.[Medline]
30. Drukker K, Horsch K, Giger ML. Multimodality computerized diagnosis of breast lesions using mammography and sonography. Acad Radiol 2005;12:970–979.[Medline]
31. Drukker K, Giger ML, Metz CE. Robustness of computerized lesion detection and classification scheme across different breast US platforms. Radiology 2005;237:834–840.[Abstract/Free Full Text]
32. Moon WK, Chang RF, Chen CJ, Chen DR, Chen WL. Solid breast masses: classification with computer-aided analysis of continuous US images obtained with probe compression. Radiology 2005;236:458–464.[Abstract/Free Full Text]
33. Chen DR, Chang RF, Chen CJ, et al. Classification of breast ultrasound images using fractal feature. Clin Imaging 2005;29:235–245.[Medline]
34. Hong AS, Rosen EL, Soo MS, Baker JA. BI-RADS for sonography: positive and negative predictive values of sonographic features. AJR Am J Roentgenol 2005;184:1260–1265.[Abstract/Free Full Text]
35. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283–298.[Medline]
36. Metz C. Evaluation of CAD methods. In: Doi K, MacMahon H, Giger ML, Hoffmann KR, eds. Computer-aided diagnosis in medical imaging. Amsterdam, the Netherlands: Elsevier Science, 1998; 543–554.
37. Zhou XH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York, NY: Wiley, 2002.

38. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 1996;201:745–750.[Abstract]
39. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York, NY: Chapman & Hall, 1993.

The Effect of Data Set Size on Computer-Aided Diagnosis of Breast Cancer: Comparing Decision Fusion to a Linear Discriminant

Jonathan L. Jesneck[1,2], Loren W. Nolte[1,3], Jay A. Baker[2], Joseph Y. Lo[1,2]

[1] Department of Biomedical Engineering, Duke University, Durham, NC 27708
[2] Duke Advanced Imaging Laboratories, Department of Radiology,
2424 Erwin Road, Suite 302, Durham, NC 27705
[3] Department of Electrical and Computer Engineering, Duke University, Durham, NC 27705

ABSTRACT

Data sets with relatively few observations (cases) in medical research are common, especially if the data are expensive or difficult to collect. Such small sample sizes usually do not provide enough information for computer models to learn data patterns well enough for good prediction and generalization. As a model that may be able to maintain good classification performance in the presence of limited data, we used decision fusion. In this study, we investigated the effect of sample size on the generalization ability of both linear discriminant analysis (LDA) and decision fusion. Subsets of large data sets were selected by a bootstrap sampling method, which allowed us to estimate the mean and standard deviation of the classification performance as a function of data set size. We applied the models to two breast cancer data sets and compared the models using receiver operating characteristic (ROC) analysis. For the more challenging calcification data set, decision fusion reached its maximum classification performance of AUC = 0.80±0.04 at 50 samples and pAUC = 0.34±0.05 at 100 samples. The LDA reached a lower performance and required many more cases, with a maximum of AUC = 0.68±0.04 and pAUC = 0.12±0.05 at 450 samples. For the mass data set, the two classifiers had more similar performance, with AUC = 0.92±0.02 and pAUC = 0.48±0.02 at 50 samples for decision fusion and AUC = 0.92±0.03 and pAUC = 0.55±0.04 at 500 samples for the LDA.

Keywords: Decision Fusion, Computer-Aided Diagnosis, Sample Size, Receiver Operating Characteristic (ROC) Curve, Classification, Breast Cancer

1. INTRODUCTION

Many medical data sets are difficult and expensive to collect, often resulting in limited data set size. A small number of cases usually precludes accurate predictive modeling. Early modeling offers many advantages, such as earlier identification of data collection problems, of unsatisfactory patient sampling, of expensive but uninformative features, and perhaps earlier discovery of flaws in the scientific experiment design. Many medical experiments expose subjects to possibly avoidable risk that could be detected by better and earlier modeling.

The amount of available data affects each model differently. Model complexity tends to produce a tradeoff between modeling power and generalization; simpler models may be more robust to noise in the data but may not be able to capture the full complexity of the data's patterns, whereas more complicated models may model the patterns better but are more susceptible to overfitting. In addition to the number of samples available, the ratio of number of features to number of samples can also affect classifier performance. Many classical models tend to overtrain on data sets with few samples and many features. This overtraining effect becomes more pronounced with smaller sample size.

In this study, we investigated the effect of sample size on the generalization ability of two computer-aided diagnosis (CADx) models. The first model was linear discriminant analysis (LDA), a common CADx model for breast cancer data. The second model was a decision-fusion method that has shown promise for small, noisy data sets[1]. Our decision-fusion technique offers the significant advantage that it can reduce the dimensionality of the feature space

of the classification problem by assigning a classifier to each feature separately. Considering only one feature at a time greatly reduces the complexity of the problem by avoiding the need to estimate multidimensional probability density functions (PDFs) of the feature space. Accurately estimating multidimensional PDFs likely requires many more observations than a typical medical data set contains[2]. Considering only one-dimensional PDFs may allow the decision-fusion technique to reach asymptotic testing performance using many fewer cases than other classifiers require.

Other benefits of decision fusion are that it is robust in noisy data[3], is not overly sensitive to the likelihood ratio threshold values[4], and can handle missing data values[5]. Our decision-fusion technique can also be tuned to optimize arbitrary performance metrics that may be more clinically relevant, unlike more traditional classification algorithms that optimize mean squared error, such as the LDA.

II. METHODS

2.1 Data

This study used two breast cancer data sets: one of mass lesions and one of calcification lesions.

The mass lesion data set is an extension of the earlier subset described by Hong, *et al.* from this research group[6]. The cases were collected between 2000 and 2005 at Duke University. The data set included 803 lesions, of which 296 were malignant and 507 were benign, and 389 were palpable and 414 nonpalpable. The patient ages ranged from 17 to 87 years, with a median age of 50 years. Patients underwent both mammography and sonography, and outcome was determined through definitive histopathological diagnosis. One of three dedicated breast radiologists with 6-11 years of experience described each lesion using Breast Imaging Reporting and Data System (BI-RADS™, American College of Radiology, Reston, VA)[7] mammography, BI-RADS sonography, and Stavros sonography descriptors[6]. Of the total 38 features, 13 were mammographic, 22 were sonographic, and 3 were patient history features.

Second, we used a calcification data set that consisted of 1508 mammogram microcalcification lesions from the Digital Database for Screening Mammography (DDSM)[8], which is publicly available. The outcomes were verified by histopathological diagnosis and follow-up for certain benign cases, yielding 811 benign and 697 malignant calcification lesions. The feature groups were 13 computer-extracted calcification cluster morphological features, 91 computer-extracted texture features of the lesion background anatomy, 2 radiologist-interpreted findings, 2 radiologist-extracted features from the BI-RADS lexicon and patient age. In total, calcification data C set had 109 features and a sample-to-feature ratio of approximately 14:1. Each mammogram was digitized with a resolution of either 43.5 microns (Howtek 960 or MultiRad850 digitizer) or 50 microns (Lumisys 200 Laser digitizer). We used a 512x512 pixel ROI centered on the centroid of each lesion (using lesion outlines drawn by the DDSM radiologists) for image processing and for generating the computer-extracted features. We extracted morphological and texture (spatial gray level dependence matrix) features, which were shown to be useful in previous studies of CADx such as by Chan, *et al.*[9].

2.1 Decision Fusion

For the decision-fusion classifier, histograms of each feature were constructed as an estimate of the probability density in order to construct an empirical likelihood ratio for that feature. Then, a binary decision was made by comparing the likelihood ratio value to a given threshold, which in turn determined the sensitivity and specificity of the decision. Finally, the decision fusion theory allowed the individual binary decisions to be combined optimally to produce one final binary decision.

First, each feature was considered separately and classified by a likelihood ratio classifier. According to decision theory, the likelihood ratio is the optimal detector to determine the presence or absence of a signal in noise[10]. The null hypothesis (H_0) was that the signal is not present in the noisy features, while the alternative hypothesis (H_1) was that the signal is present.

$$H_0 : X = N$$
$$H_1 : X = S + N \tag{1}$$

The likelihood ratio is the probability of the features under the malignant case divided by the probability of the features under the benign case:

$$\lambda(X) = \frac{P(X \mid H_1)}{P(X \mid H_0)}, \tag{2}$$

where $p(X|H_1)$ is the PDF of the observation data X given that the signal is present, and $p(X|H_0)$ is the PDF of the data X given that the signal is not present. The likelihood ratio is optimal under the assumption that the PDFs accurately reflect the true densities. For classification, we can apply a threshold value, τ, to the likelihood ratio to produce a binary decision, u, about the presence of the signal.

$$u = \begin{cases} 1 & \text{if } \lambda \geq \tau \\ 0 & \text{if } \lambda < \tau \end{cases} \tag{3}$$

Since we assigned a separate likelihood ratio classifier to each of *p* features, we applied a separate threshold to each classifier's output value to produce *p* binary decisions. A genetic algorithm searched over the joint set of thresholds in order to maximize the classification performance of the fused binary decisions. The genetic algorithm search time was capped at 30 generations for this study due to computational cost.

Decision-fusion theory describes how to combine local binary decisions optimally to determine the presence or absence of a signal in noise[11-15]. The decision fuser optimally fuses all the local decisions according to the operating points on the receiver operating characteristic (ROC) curve at which the local decisions were made. Assuming statistically independent decisions, the likelihood ratio of the fused classifier is a product over the "yes, signal present" ($u_i = 1$) decisions multiplied by a similar product over the "no, signal absent" ($u_i = 0$) decisions.

$$\lambda_{fused}(u_1,\ldots,u_p) = \prod_{\substack{i \\ u=1}} \frac{Pd_i}{Pf_i} \prod_{\substack{i \\ u=0}} \frac{1-Pd_i}{1-Pf_i}, \tag{4}$$

where Pd_i is the probability of detection or sensitivity, and Pf_i is the probability of false detection, or (1-specificity), for the i^{th} local decision. The ROC curve can be computed from the unique likelihood-ratio values of the fused classifier as shown in Equation (5).

$$\begin{aligned} Pd_{fused}(j) &= \sum_{i=j}^{p} P(\lambda_{fused,i} \mid H_1), \quad j = 0,\ldots,p \\ Pf_{fused}(j) &= \sum_{i=j}^{p} P(\lambda_{fused,i} \mid H_0), \quad j = 0,\ldots,p \end{aligned} \tag{5}$$

2.2 Linear Discriminant Analysis

The baseline classifier was linear discriminant analysis (LDA), which served as a benchmark for the linear separability of the data set.

2.3 Sampling and Validation

In order to study the effect of sample size on the classifiers' performances, we randomly selected subsets of the data sets. We varied the number of selected cases from 50 to 500, which covers typical data set sizes in preliminary CADx research. Ten random draws of each data subset size were drawn to assess selection effects. On each subset, both classifiers were trained and validated using 10-fold cross-validation. For each sample size such as 100 cases, classifiers were developed using ten bootstrap samples of that number of cases, which allowed the calculation of the mean AUC and pAUC values along with their standard deviations.

2.4 Classifier Comparison

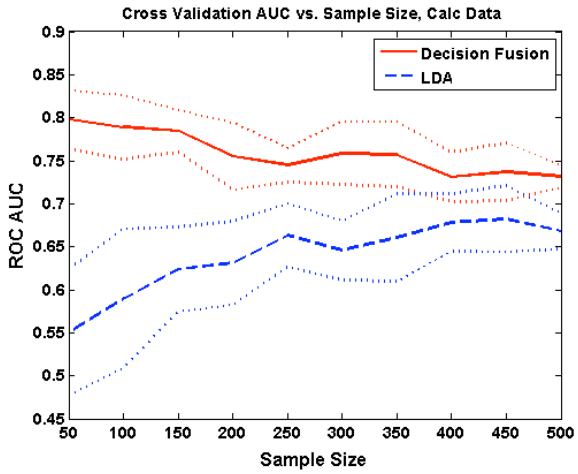
Each classifier was evaluated using ROC analysis. Two clinically interesting summary metrics of the ROC curve were used: the area under the curve (AUC) and the normalized partial area of the curve (pAUC), which is measured above sensitivity of Pd = 0.9.

III. RESULTS

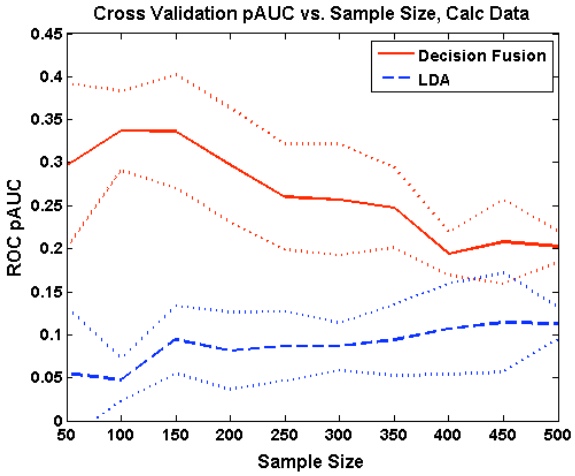
Figure 1 plots the classification performance against the number of cases. The classifiers' performances were scored both by ROC AUC (Fig. 1a and 1c) and pAUC (1b and 1d).

On the calcification data (Fig. 1a and 1b) decision fusion achieved a maximum of AUC = 0.80±0.04 at 50 samples and pAUC = 0.34±0.05 at 100 samples. The LDA had a lesser performance, with AUC = 0.68±0.04 and pAUC = 0.12±0.05 at 450 samples. The LDA had the expected testing trend of slowly increasing performance with increasing sample size, but decision fusion showed the opposite trend. Perhaps inadequately trained, decision fusion decreased with sample size both in AUC and pAUC. Note that all of these are validation results from k-fold cross-validation, which normally should minimize effects of training bias.

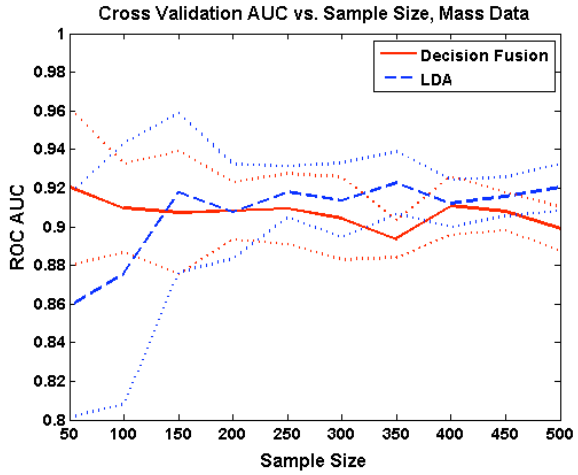
For the mass lesion data (Fig. 1b and 1d), the two classifiers' performances had more similar trends. Decision fusion reached a maximum of AUC = 0.92±0.02 and pAUC = 0.48±0.02 at 50 samples, and the LDA reached AUC = 0.92±0.03 and pAUC = 0.55±0.04 at 500 samples. No significant performance differences between the classifiers were seen in sample sizes greater than 100. For very small data sets of 50 cases, decision fusion outperformed the LDA. In both data sets, decision fusion approached its final AUC value with many fewer cases than the LDA required. All plots except Fig. 1b showed that decision fusion had a smaller slope than the LDA.



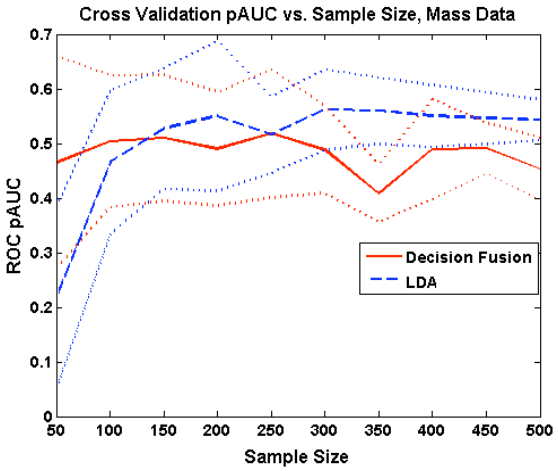
(a) AUC vs. Sample Size, Calcification Data



(b) pAUC vs. Sample Size, Calcification Data



(c) AUC vs. Sample Size, Mass Data



(d) pAUC vs. Sample Size, Mass Data

Figure 1: Classifier performance vs. Sample Size

Decision fusion significantly outperformed the LDA on the calcification data set. The performance difference was greatest for small data sets. However, on the larger data sets, the performance gap narrowed to 0.06. In part (b), decision fusion achieved pAUC = 0.34±0.05 at 100 samples and then fell to pAUC = 0.2±0.02 at 500 samples. Although the two classifiers had very similar performance on the mass data set, decision fusion still outperformed the LDA for very small sample sizes.

IV. DISCUSSION

Decision fusion had its biggest classification performance gain over the LDA on the noisier, more nonlinear data set, the calcification data set. On the mass data set, both the LDA and decision fusion performed very similarly for data sets larger than 50 samples. On very small data sets of 50 samples, which are common among initial CADx studies, decision fusion outperformed the LDA. For the mass data set at least, a particular strength of the decision-fusion algorithm is that it is able to estimate asymptotic testing performance with many fewer cases than other classifiers require. Figure 1 shows that decision fusion was able to achieve approximately the same testing performance with 50 cases as with 500 cases.

The general downward slope of the decision fusion curves for the calcification data set may be due to inadequate training. For computational convenience, we limited the genetic algorithm's search time to only 30 generations. Whereas 30 generations were adequate for small data sets smaller than 150 cases, larger data sets required more genetic algorithm generations for complete optimization. A much longer run of 3000 generations on all available 1508 cases in the calcification lesion data set improved decision fusion's performance under 100-fold cross-validation to AUC = 0.85±0.01 and pAUC = 0.28±0.03, which exceeded the performance for all data points shown in Fig. 1a and 1b. A similar more thorough optimization on all available 803 cases in the mass data set allowed decision fusion to reach AUC = 0.94±0.01 and pAUC = 0.63±0.07, which likewise also exceeded the performances in Fig. 1c and 1d.

The improvements were usually significant for the more challenging calcification data set, but not for the mass data set. Such a statement may not reflect the full diversity of these data sets, which differ in many respects, including linear separability, numbers of cases, numbers and types of features, and feature correlations. Future work will explore the contribution of such factors using controlled simulation data sets in order to understand the full potential and limitations of the decision-fusion technique.

ACKNOWLEDGEMENTS

This work was supported by US Army Breast Cancer Research Program W81XWH-05-1-0292 and DAMD17-02-1-0373, and NIH/NCI R01 CA95061 and R21 CA93461. I would also like to thank Brian Harrawood for the ROC bootstrap code, Anna Bilska-Wolak, Ph.D., and Georgia Tourassi, Ph.D., for insightful discussions, and Andrea Hong, M.D., Jennifer Nicholas, M.D., Priscilla Chyn, and Susan Lim for data collection.

REFERENCES

1. Jesneck JL, Lo JY, Baker JA. "A computer aid for diagnosis of breast mass lesions using both mammographic and sonographic descriptors" Radiology (submitted January 2006).
2. Hastie T, R. T, Friedman JH. *The Elements of Statistical Learning*, Springer, 2001.
3. Niu R, Varshney PK, Moore M, Klamer D. "Decision fusion in a wireless sensor network with a large number of sensors". International Society of Information Fusion, Fairborn, OH 45324, United States, 2004; 21.
4. Zhu M, Ding S, Brooks RR, Wu Q, Rao NSV, Iyengar SS. "Fusion of threshold rules for target detection in sensor networks". ACM Transactions on Sensors Networks (submitted for publication).
5. Bilska-Wolak AO, Floyd CE, Jr. "Tolerance to missing data using a likelihood ratio based classifier for computer-aided classification of breast cancer". Phys Med Biol 2004; **49**:4219-4237.

6. Hong AS, Rosen EL, Soo MS, Baker JA. "BI-RADS for Sonography: Positive and Negative Predictive Values of Sonographic Features". Am J Roentgenol 2005; **184**:1260-1265.
7. American College of Radiology, *Breast Imaging - Reporting and Data System (BI-RADS)* 3rd ed., Reston, VA, American College of Radiology, 1998.
8. Heath M, Bowyer KW, Kopans D. "Current status of the Digital Database for Screening Mammography". In: *Digital Mammography*, Karssemeijer N, Thijssen M, Hendriks J, eds.: Kluwer Academic Publishers, p. 457-460, 1998.
9. Chan HP, Sahiner B, Lam KL, et al. "Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces". Medical Physics 1998; **25**:2007-2019.
10. VanTrees HL. *Detection, Estimation, and Modulation Theory (Part I)*. John Wiley & Sons, New York, 1968.
11. Tenney RR, Sandell NR, Jr. "Detection with Distributed Sensors". *Proc IEEE Conf Incl Symp Adapt Processes*, 1980; **1**:433.
12. Chair Z, Varshney PK. "Optimal data fusion in multiple sensor detecton systems". *IEEE Transactions on Aerospace and Electronic Systems* 1986; AES-**22**:98.
13. Reibman AR, Nolte LW. "Optimal detection and performance of distributed sensor systems". *IEEE Transactions on Aerospace and Electronic Systems* 1987; AES-**23**:24.
14. Dasarathy BV. "Decision fusion strategies in multisensor environments". *IEEE Transactions on Systems, Man and Cybernetics* 1991; **21**:1140.
15. Liao Y. *Distributed decision fusion in signal detection -- a robust approach*. Ph.D. Thesis, Duke Univeristy, 2005.

A Bayesian method to estimate the minimum sample size for decision fusion

Jonathan L. Jesneck[1,2,3,4], B.S.E., Loren W. Nolte[1,5], Ph.D., Georgia D. Tourassi[1,3,6], Ph.D., Joseph Y. Lo[1,3,6], Ph.D.

1. Department of Biomedical Engineering
2. Institute of Statistics and Decision Sciences
3. Duke Advanced Imaging Labs, Department of Radiology
4. Institute of Genome Sciences and Policy
5. Department of Electrical and Computer Engineering
6. Medical Physics Graduate Program

February 2007

Intended journal: Medical Decision Making (impact factor 1.822)

Abstract

Background

To serve as useful and reliable medical tools, computational models and computer-aided diagnosis (CADx) systems must be properly trained. Training models on complex medical data often requires numerous training samples, which may be expensive to collect. To minimize the data collection costs and to ensure that the models generalize well to future cases, it is important to estimate the minimum number of samples needed for a particular modeling and classification task.

Method

This study focuses on the decision-fusion classifier, which optimally combines input binary decisions. Here we present a Bayesian approach to estimate the minimum number of training samples at which decision fusion reaches its asymptotic testing performance. To demonstrate the actual performance of the described Bayesian theory, we performed a large series of Monte Carlo simulation studies, varying parameters such as feature count and feature strength. The simulation results were then used to make sample size inferences on real medical data sets.

Results

Minimum training set size decreased with both signal strength (signal-to-noise ratio, SNR) and the number of features at a constant SNR per feature. Minimum sample size increased when a constant whole-dataset SNR was diluted across various numbers of features.

Conclusions

For a decision-fusion classification task, this method uses signal strength and feature count to estimate the minimum sample size. Having enough training samples allows the classifier to be adequately trained to generalize optimally on future cases. To help extend this simulation study for application on real medical data, readers are provided with sampling guidelines to estimate the minimum training set size for their own data sets.

Key Words

decision fusion; sample size; Bayesian estimation; generalization; ROC curve; computer-aided diagnosis (CADx)

End of Abstract

I. Introduction

Computational models and computer-aided diagnosis (CADx) systems are becoming increasingly important in medicine. In order for these models to serve as a useful medical tool, they must be properly trained and must generalize well to new cases. Good generalization depends heavily upon proper training, which requires an adequate number of training samples. However, training cases are often difficult to acquire; collecting medical data is often time-consuming and expensive. To make best use of limited resources and to allow computational models to be properly trained, it is important to estimate the minimum required size of a data set.

Many studies have shown that small data sets adversely affect modeling efforts. Fukunaga and Hayes [1] and Raudys and Jain [2] have conducted reviews of the large literature of finite-

sample effects. Pertaining specifically to finite-sample effects on CADx, Chan *et al.* [3, 4] showed how a small sample size introduces bias and degrades the CADx classifier performance. This effect has been explored in depth for linear classifiers [5, 6] and artificial neural networks (ANNs) [7-9]. Finite sample size is especially important when classifier training is performed in conjunction with classifier design and feature selection [6, 8, 10].

The finite-sample effect confounds traditional classifiers because they have only a few samples from which to estimate the structure of a complicated data set. Estimating multidimensional probability density functions (PDFs) is a difficult problem in statistical learning. One method for tackling this problem is to consider only one dimension at a time. With limited observations, it is far easier to estimate a one-dimensional PDF than a high-dimensional PDF.

One classifier designed to meet this challenge is decision fusion. The decision fusion algorithm operates by the following two steps: (1) Local classifiers use feature subsets to generate initial binary decisions, and (2) With decision-fusion theory we then combine these binary decisions optimally. The algorithm is described in detail in previous work [11]. One of decision fusion's main advantages is that it avoids the problem of having to estimate multidimensional probability distribution functions (PDFs). Our decision fusion technique reduces the dimensionality of the classification problem's feature space by initially considering only one feature at a time. By estimating only the one-dimensional distribution of each feature separately, decision fusion is able to capture underlying trends in the data by using fewer training samples than many multidimensional modeling techniques.

The purpose of this study was to identify the minimum sample size for a decision fusion classification. The minimum sample size was defined as the minimum number of training samples that decision fusion required to achieve its asymptotic classification performance. We

developed a Bayesian method to model the uncertainty from the finite-sample effect and explore this uncertainty's effect on decision fusion's classification performance. This method's performance trends were shown in a large series of Monte Carlo simulation runs.

II. Methods

A. Decision fusion

Stemming from the field of distributed detection, decision fusion has a growing literature. In early work, Tenney and Sandel [12] optimized the local processors and kept fixed the fusion processor, whereas Chair and Varshney [13] fixed the local processors and optimized the fusion processor. Reibman and Nolte [14] optimized the local and fusion processors simultaneously and derived the overall optimum fusion design. Dasarathy [15] summarized this previous work. Jesneck *et al.* [11] used genetic algorithms to optimize of the decision fusion algorithm into arbitrarily high dimensions.

The goal of decision fusion is to detect the presence or absence of a signal in noise [16]. Signal detection is formulated in decision fusion theory by optimally combining local binary decisions [11-15, 17]. Although the local binary decisions can come from any arbitrary source, previous work [11, 18, 19] demonstrated the benefit of using a likelihood-ratio classifier. The likelihood ratio is probability of the feature assuming the signal is present in noise ($H_1 : X = S + N$) divided by the probability of the feature assuming the signal is absent ($H_0 : X = N$):

$$\lambda_{feature}(X) = \frac{P(X \mid H_1)}{P(X \mid H_0)} \tag{1}$$

where S is the signal, N is noise, $P(X \mid H_1)$ is the PDF of the observation data X given that the signal is present, and $P(X \mid H_0)$ is the PDF of the data X given that the signal is not present.

Once the classifier output value $\lambda_{feature}$ has been calculated, it can then be subjected to a threshold τ to make a binary decision u.

$$u = \begin{cases} 1 & \text{if } \lambda_{feature} \geq \tau \\ 0 & \text{if } \lambda_{feature} < \tau \end{cases} \tag{2}$$

Any particular threshold determines a corresponding operating point on the local classifier's receiver operating characteristic (ROC) curve. Each ROC operating point consists of a pair of probabilities (Pd, Pf), where Pd is the probability of detection, and Pf is the probability of false alarm, at which a local binary decision is made.

Our decision fusion algorithm applies a separate likelihood-ratio classifier to each feature. The local classifier receives an input feature value, X, and produces an output value, $\lambda_{feature}$, which is compared to a threshold τ to yield a binary decision u. The likelihood ratio of a single binary decision u can be written in a simple ratio form:

$$\lambda_{decision}(u) = \frac{P(u \mid H_1)}{P(u \mid H_0)} = \begin{cases} \dfrac{Pd}{Pf} & \text{if } u = 1 \\ \dfrac{1 - Pd}{1 - Pf} & \text{if } u = 0 \end{cases} \tag{3}$$

The set of binary decisions and their ROC operating points are then fused into a likelihood-ratio value, which takes the form of a product over the "yes" decisions ($u_i = 1$) and the "no" decisions ($u_i = 0$):

$$\lambda_{fusion}(u_1,...,u_p) = \prod_{i=1}^{p} \lambda_{decision}(u_i)$$
$$= \prod_{i=1}^{p} \frac{P(u_i \mid H_1)}{P(u_i \mid H_0)} \tag{4}$$
$$= \prod_{i=1}^{p} \left(\frac{Pd_i}{Pf_i}\right)^{u_i} \left(\frac{1-Pd_i}{1-Pf_i}\right)^{1-u_i}$$

where Pd_i and Pf_i are the probabilities of detection and false alarm, respectively, for the i[th] local binary decision, u_i.

With this product form of the decision fusion likelihood, we assume that the local decisions are statistically independent [11]. While we could in fact construct an optimal correlated decision fusion processor with known decision correlations [20], correctly estimating the decision correlations would require a large number of training samples. Limited to few observed samples, therefore, we make the decision independence assumption. Although the decision independence assumption appears to be very strong, its application in decision fusion often does not substantially lower classification performance in practice [17]. Note that we assume statistical independence only of the binary decisions, not of the sensitivity, false-positive rate, or even the features on which the local decisions were made.

To measure classification performance, the decision-fusion likelihood-ratio value λ_{fusion} can be used as the output value of the decision fusion algorithm. By applying thresholds to this output value, we can use ROC analysis to measure the algorithm's classification performance. In addition to a qualitative assessment of the ROC curve's shape, we can quantify performance by using figures of merit, such as the area under the ROC curve (AUC).

B. Uncertainty about the values of the local ROC operating points

The optimality of decision fusion is guaranteed only if we know the ROC operating points exactly. These values are easily calculated on training data; a threshold τ applied to the observed classifier output value $\lambda_{feature}$ yields the operating point values $\left(Pd_{train}, Pf_{train}\right)\Big|_{\tau}$, as shown Equation 2. But in order for decision fusion to achieve optimal generalization performance, we must also be able to determine the ROC operating points for future cases. However, for testing data or future samples this calculation is impossible because we cannot explicitly determine which ROC operating points are determined by that same threshold τ. When applied to new samples, this threshold will yield different operating point values $\left(Pd_{test}, Pf_{test}\right)\Big|_{\tau}$. Large differences between the training and testing ROC operating points, $\left(Pd_{train} - Pd_{test}, Pf_{train} - Pf_{test}\right)\Big|_{\tau}$, will degrade the decision fusion's performance on future observations. But small differences will allow decision fusion to generalize at near-optimal levels. Statistical learning theory states that such a training-testing difference shrinks with more training samples, because both observed training and testing values converge to asymptotic values determined by the data's underlying probability distribution [21].

Therefore we can use a statistical framework to estimate how many training samples are required for good generalization on future cases. In order to estimate the minimum number of samples required for decision fusion to reach its asymptotic classification performance, we can model the uncertainty of the Pd and Pf estimates. By choosing different thresholds, we can estimate the uncertainty of each operating point on the ROC curve. Figure 1 shows a local classifier's ROC curve with its uncertainty band. An arbitrary threshold determines a particular operating point, shown by the dot. For this study we chose the threshold $\tau = 1$. The vertical and horizontal arrows in Figure 1 represent the ROC operating point's uncertainty. The following statistical methodology describes how to model this uncertainty.

C. Estimating the $\left(Pd, Pf\right)\Big|_{\tau}$ uncertainty

To estimate the uncertainty of the ROC operating points, one could use any of the following three estimation methods: ROC parametric fitting, sampling of a fixed ROC curve, and sampling of the input data.

In parametric modeling of the ROC curve, a common assumption is the binormal assumption: the data points from class 0 (hypothesis H_0) and class 1 (hypothesis H_1) have normal distributions [22-24]. Under this assumption, the ROC parameters are fit. Along with point estimates for the curve's parameters, the fit also gives us the parameters' variances. With these variances we can model the uncertainty of the ROC operating points and estimate the minimum sample size [25]. This study did not use this ROC modeling technique in order to avoid the potential limitations of the binormal assumption.

In nonparametric bootstrap sampling, we use an empirical ROC curve. We can perform bootstrap sampling over the set of values of the output decision variable, which generates a set of bootstrapped ROC curves [26]. From generated ROC curves, we can calculate the confidence bands along the ROC curve. For any particular operating point, this method can calculate the Pd and Pf variances. Note that for this method the classifier is run only once, producing a single set of values of it output decision variable. Sampling is then performed on these output values to generate many similar ROC curves. This study did not use this technique in order to avoid possible biases due to single draws of small samples that were not representative of their underlying distributions.

For a nonparametric method that is more computationally expensive but also potentially less biased, we could sample the input data points. Whereas the above method performs sampling after a single run of the classifier, this method runs the classifier for each sampling draw. With

each data sample, the classifier runs and generates an ROC curve. It is this second nonparametric sampling method that we have used in this study.

D. A Bayesian method to integrate over the $\left(Pd_{train}, Pf_{train}\right)\Big|_{\tau}$ uncertainty

Once we have estimated the ranges of the ROC operating points, we can pass this information to the decision fusion algorithm by constructing a Bayesian model. In a Bayesian setting, we define an *a priori* distribution of the ROC operating points. Although any form for the distribution can be used, we chose a small uniform distribution centered on the training estimates $\left(Pd_{train}, Pf_{train}\right)\Big|_{\tau}$ [17].

$$\begin{aligned} Pd\big|_{\tau} &\sim U\left(Pd_{train}\big|_{\tau} - \delta_{Pd}, Pd_{train}\big|_{\tau} + \delta_{Pd}\right) \\ Pf\big|_{\tau} &\sim U\left(Pf_{train}\big|_{\tau} - \delta_{Pf}, Pf_{train}\big|_{\tau} + \delta_{Pf}\right) \end{aligned} \tag{5}$$

Here $2\delta_{Pd}$ and $2\delta_{Pf}$ are the distribution widths of Pd and Pf, respectively. The *a priori* distribution of $\left(Pd, Pf\right)\Big|_{\tau}$ allowed us to integrate over the $\left(Pd, Pf\right)\Big|_{\tau}$ uncertainties in order to get a marginal estimate of the decision-fusion likelihood-ratio value.

$$\lambda_{fused}\Big|_{\tau} = \int_{Pd_1}\int_{Pf_1}\cdots\int_{Pd_p}\int_{Pf_p}\left(\lambda_{fused}\Big|Pd, Pf\right)\Big|_{\tau} \; d(Pd_1)d(Pf_1)\ldots d(Pd_p)d(Pf_p) \tag{6}$$

Since this integral has no known analytical solution, it was approximated with a Monte Carlo simulation study.

E. Monte Carlo simulation

Our Monte Carlo simulation consisted of three main steps: (1) Generate input data, (2) Train the decision fusion classifier to the observed data and model the uncertainty of the ROC operating points, and (3) Apply the trained decision fusion classifier to the testing data. These three steps were performed under various conditions, so as to observe trends in decision fusion's classification performance. The goal was to identify the minimum number of training observations needed for the decision fusion algorithm to achieve its asymptotic classification performance.

The first step of the Monte Carlo simulation was to generate the input data X. We drew sample X values from normal distributions. Both training and testing data sets were drawn for each Monte Carlo sampling iteration. These samplings produced a large set of ROC curves. On these ROC curves we applied the threshold $\tau = 1$ as in Equation 2 and calculated the 5% and 95% quantiles for the ROC operating points.

For the second step, we used the 5% and 95% percentiles to set the domains of the *a priori* distributions of the ROC operating points $(Pd, Pf)\big|_{\tau}$. Then we drew sample ROC operating points from their respective distributions (Equation 5). For each draw, we computed the decision fusion likelihood-ratio value λ_{fusion} (Equation 4). Next, we averaged the computed likelihood-ratio values over all the drawn ROC operating points to yield a marginal estimate of the likelihood-ratio value:

$$\hat{\lambda}_{fused} = \frac{1}{pM} \sum_{Pd_1} \sum_{Pf_1} \cdots \sum_{Pd_p} \sum_{Pf_p} \lambda_{fused} \big| Pd, Pf \tag{7}$$

Here p is the number of features and M is the number of draws in the Monte Carlo simulation. We ran the simulation for approximately 500,000 sampling iterations.

In the third step, the trained decision fusion models were applied to independent testing data. Classification performance was measured with ROC analysis.

F. Investigating decision fusion's asymptotic behavior

Using the above Bayesian method, we investigated the asymptotic classification performance of the decision fusion algorithm. The goal was to determine the minimum number of training samples at which the asymptotic value was reached. The training sets had varying numbers of observations, from 10 to 1000, whereas testing sets always comprised at least 1000

observations. The training and testing sets were independent. As the sample size varied, the performance trends were analyzed under various data set conditions and algorithmic parameter settings. Over the simulated data sets, we varied the features' signal-to-noise ratios (SNRs) and the number of features, which determined the number local binary decisions to fuse. We varied the degree of overlap between the distributions of the two classes, which created various signal strengths, with signal-to-noise ratios (SNRs) ranging from 0.05 to 5.0 per feature.

To measure the effect of diluting the signal across features, we also held constant the signal strength of the entire data set, but we spread the signal evenly across various numbers of features. For example, we chose a whole data set SNR = 10, which we divided into 2 features at SNR = 5.0 each, 5 features at SNR = 0.4 each, 10 features at SNR = 0.2 each, and 20 features at SNR = 0.1 each. We identified the number of training samples at which the asymptote was reached by asymptote was reached when the AUC was within 0.01 of the asymptote: $\left|AUC_n - AUC_\infty\right| < 0.01$ where AUC_n is the testing AUC for *n* training samples, and AUC_∞ is the asymptotic testing performance.

In order to demonstrate decision fusion's performance on data sets of many very weak features, we spread the whole data set SNR = 10 across 200 features at SNR = 0.05 each. At such a low signal strength, each feature's signal was deeply buried in noise. We ran both the Bayesian decision fusion algorithm and linear discriminant analysis (LDA) on this weak data set, for various numbers of training samples.

III. Results

Table 1 lists the asymptotic testing performance levels and the minimum sample size for various experimental conditions. The Monte Carlo simulations showed that more training samples increased the expected value and decreased the variance of the testing metric AUC_{test}. Decision

fusion achieved its highest testing performance with more local binary decisions coming from stronger features (higher SNRs).

Although the Monte Carlo simulation was run many times for different settings of various variables, for clarity we focus here on one representative algorithmic setting: the fusion of two decisions, coming from features with a SNR of 1.0 and created by applying the threshold $\tau = 1$. To investigate how the training set size affected decision fusion's testing performance, we varied the number of training samples from 10 to 1,000. Figure 2 shows means and ranges of the ROC operating points, both training and testing. Note that we use the notation Pd and Pf for the theoretical operating point values but TPF and FPF for their respective observed values. The drawn TPF and FPF values reached their asymptotic distributions for training sets of 500 or more samples. Figure 3 shows the operating points' estimation errors $\left(TPF_{train} - TPF_{test}\right)\Big|_{\tau=0}$ and $\left(FPF_{train} - FPF_{test}\right)\Big|_{\tau=0}$. Similarly to the operating points' values, their errors also reached their asymptotic distributions at 400 observations. These drawn operating point values created the set of decision fusion ROC curves shown in Figure 4a. As the number of training samples increased, the band of resulting ROC curves shifted from a loose band around the chance diagonal line (for 10 training samples, shown in red) to a tight band at much higher performance (for 1000 training samples, shown in magenta). The areas under these ROC curves are shown at the red line in Figure 5c. The bold, solid red line delineates the mean of the AUC values, and the dotted red lines mark the 5% and 95% percentile ranges. Agreeing with Figures 2, 3, and 4a, this figure shows that at approximately 400 training samples decision fusion reached its asymptotic testing performance: AUC = 0.75 (with a 90% confidence interval of 0.74 to 0.76).

Figures 4 and 5 also show the testing performance for other algorithmic settings, such as other numbers of decisions to fuse and stronger or weaker features. In Figure 5 the input features varied from weak with SNR = 0.1 to stronger with SNR = 1.0. For the very weak features (Figure

5a), none of the decision fusion curves attained their asymptotes by 1000 training samples. Stronger features (Figures 5b and 5c) allowed the decision fusion algorithm to reach its asymptotic values.

Figure 6 shows the testing performance with the whole data set signal held constant at SNR = 10. In different algorithm runs, the signal was split over 2, 5, 10, and 20 decisions. For each scenario, decision fusion achieved the same asymptotic testing performance, $AUC_{\infty} = 0.95 \pm 0.01$. This asymptotic testing performance was reached at 50 training samples for 2 decisions, at 90 training samples for 5 decisions, at 300 training samples for 10 decisions, and at 400 training samples for 20 decisions.

Comparing the testing performances of decision fusion and LDA, Figure 7 shows the classifiers' performance trends on a very weak data set of 200 features, each with SNR = 0.05. The mean testing AUC values appear in bold, and the thin dotted curves indincate the 5% to 95% quantile bands. Both classifiers started at near-chance, poor discrimination for very small sample sizes. Decision fusion outperformed LDA at 100 training samples ($p < 10^{-6}$) and even at 20 training samples ($p < 10^{-3}$), although, with so few training cases, both classifiers performances were only marginally better than chance guessing ($AUC = 0.51$ for LDA and $AUC = 0.54$ for decision fusion).

IV. Discussion and Conclusion

Because of the high cost of collecting many medical data sets, it is important to estimate the minimum sample size needed for a scientific study. Although calculating this estimate is difficult for many machine learning algorithms, it is easier for decision fusion. This study describes a Bayesian technique for estimating the minimal number of training samples at which decision fusion reaches its asymptotic testing performance. Numerous Monte Carlo simulations investigated the effect of training sample size on decision fusion's testing behavior.

The Monte Carlo simulations demonstrated certain general decision fusion performance trends. As expected, small training sets created test ROC curves that are near the chance diagonal line. Larger data sets created better-performing test ROC curves. These test ROC curves approached their asymptotic shape as the number of training observations is increased.

As expected, adding more signal energy to the data set increased the classification performance. Keeping constant the SNR per feature, the test performance increased with more local decisions to fuse. The asymptotic testing levels always rose sequentially from the task of fusing two decisions to that of fusing 100 decisions. For very small training set sizes, however, sparse-sampling effects caused the testing curves to cross; fusion of 100 decisions underperformed the fusion of fewer decisions (Figure 5). However, the performance differences here were not statistically significant. In addition to boosting classification performance, more local decisions also decreased the variances of the resulting ROC curves and their areas. The testing performance also rose with increasing signal strength of the features. Weak features caused the decision fusion algorithm to need many more training samples in order to reach its asymptotic testing performance. More training features were also needed for the Pd and Pf draws to reach their asymptotic distributions.

Holding the whole-dataset SNR constant, we spread the signal energy across various numbers of features. This signal spreading added uncertainty to the decision fusion problem. To recover from this extra uncertainty and dimensionality, decision fusion needed more training samples. However, as long as the whole-dataset SNR was held constant, decision fusion eventually reached the same asymptotic AUC performance level, as shown in Figure 6. To illustrate an extreme example of signal spreading, a whole-dataset SNR of 10 was spread across 200 features, creating a large number of very weak features. Figure 7 shows that decision fusion outperformed an LDA, even on very few training samples. Although admittedly these

performance values are poor for this extremely challenging data set, Figure 7 demonstrates decision fusion's performance benefit over more traditional classifiers. Because the decision fusion algorithm does not need to estimate multidimensional PDFs, it is an especially useful tool for weak datasets with few training samples.

The presented decision fusion algorithm was based on two important assumptions. The first assumption was that the local binary decisions were statistically independent. Although this independence assumption seems restrictive, it often does not lower the classification performance below that of the optimal decision fusion processor for correlated decisions, which requires more data to train accurately [17]. Note that we assume only the local binary decisions to be statistically independent, but not the sensitivity, false-positive rate, or even the features on which the local decisions were made. The second assumption was that the local classifiers' outputs were normally distributed. This assumption was made only for convenience in this simulation study, but it is not necessary for real data. The distributions of the local classifiers' ROC operating points can be estimated using either a parametric fit for the ROC curve or a bootstrap sampling technique.

In our previous work [11], we applied the decision fusion algorithm to two real medical data sets. We can use the sample size methods described in this study to estimate whether those medical data sets were large enough for decision fusion to achieve its asymptotic performance. Data set M consisted of breast mass features, and data set C consisted of breast calcification features. The classification goal was to distinguish benign from malignant lesions. The breast mass data set M (mass lesions) was the easier classification problem, with decision fusion yielding $AUC = 0.94 \pm 0.01$. Figure 8a shows the histogram of SNR values of the 38 features. Data set C (calcification lesions) presented a much more challenging classification problem (Figure 8b), with 110 very weak features (almost all with $SNR < 0.10$). The large number of very weak features lead decision fusion to yield $AUC = 0.85 \pm 0.01$, even with 1508 training samples. By

comparing the decision fusion performance on similar generated data sets from the simulation study, the described Bayesian methodology suggested that the breast mass data set M did in fact have enough training samples for decision fusion to reach its asymptotic performance level. The calcification data set C, however, did not have enough samples for decision fusion because of the large number of very weak features.

Although the majority of results presented here were from simulated data, it is important to explain how to use this algorithm on real medical data. Researchers can use the following steps:

1. Choose local binary classifiers for the features. Each feature should have its own separate classifier, such as a likelihood ratio classifier.
2. Pick ROC operating points at which these local classifiers will operate. Previous work [11] showed how to use a genetic algorithm to choose these local operating points in order to optimize decision fusion's classification performance.
3. Estimate the uncertainty range of the chosen ROC operating points. This estimation can be done by either the ROC curve fitting or nonparametric sampling methods, as described in section C.
4. Integrate over the operating point uncertainty using the described Bayesian approach, which is implemented by a Monte Carlo simulation. Run the decision fusion algorithm using the drawn ROC operating point values.

The source code is available on the Supplemental Materials website.

In conclusion, we have described a Bayesian technique to estimate the minimum number of training samples at which decision fusion reaches its asymptotic testing performance. Although the technique has been introduced on simulated data, the approach is general and can be applied to real medical data.

Acknowledgments

This work was supported by the U.S. Army Breast Cancer Research Program (Grant No. W81XWH-05-1-0292) and NIH/NCI R01-CA95061-01 and R01 CA101911. We thank Tobias Sing and Oliver Sander for their ROCR software to efficiently calculate large sets of ROC curves.

List of Tables

Number of decisions	Feature SNR	Asymptotic AUC_{test}	Minimum required training samples
2	0.1	0.58 (0.56, 0.59)	1000
2	1.0	0.76 (0.76, 0.77)	550
5	0.1	0.65 (0.63, 0.66)	950
5	1.0	0.89 (0.88, 0.89)	450
10	0.1	0.69 (0.67, 0.72)	700
10	1.0	0.95 (.095, 0.96)	550
100	0.1	0.86 (0.83, 0.89)	850
100	1.0	1.0 (0.99, 1.0)	70

Table 1: Asymptotic classification testing performance levels of decision fusion under various experimental conditions.

List of Figures

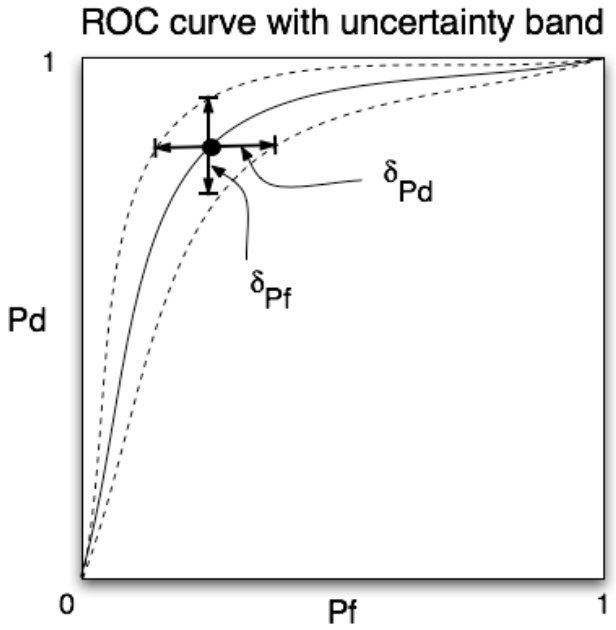
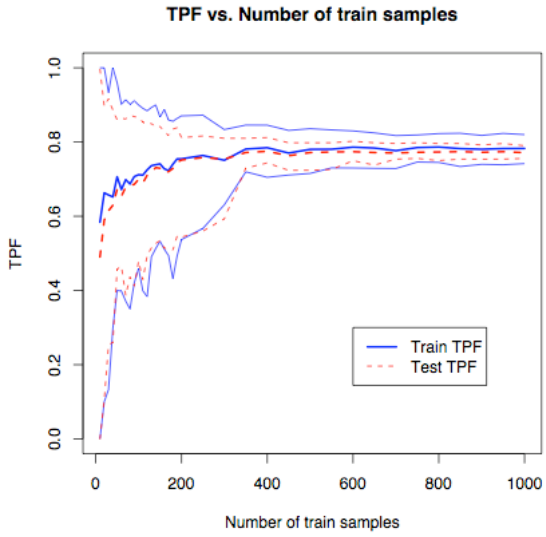
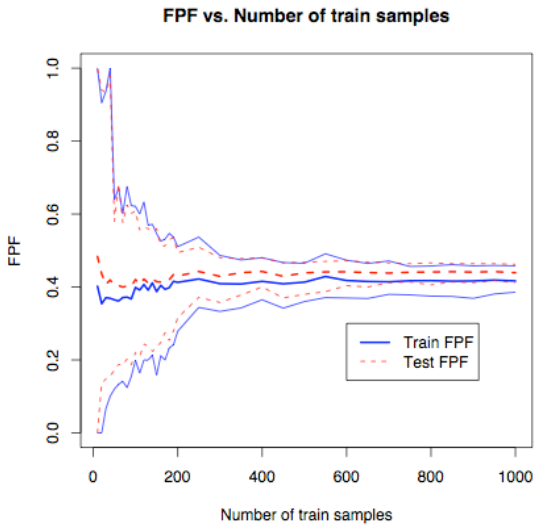


Figure 1: Diagram of uncertainty of an ROC operating point. Uncertainty is formulated as a statistical distribution of ROC curves. Solid line shows the observed ROC curve, and dotted lines form 5% and 95% percentile bands. A particular threshold determines a specific operating point, shown by the dot. Distance between bands shows ranges of the specific operating point probability of false alarm Pf horizontally and probability of detection Pd vertically. As the number of training observations is increased, leading to asymptotic classification performance, the testing ROC curve and its uncertainty band approach their stable shapes and positions.





(a) Drawn TPF values and training set size (b) Drawn FPF values and training set size

Figure 2: Monte Carlo simulations of ROC operating points determined by the threshold $\tau = 1$, for the task of fusing two local binary decisions. Each local binary decision resulted from the threshold applied to noisy feature variables with signal-to-noise ratio (SNR) of 1.0. Figure 2a shows the true positive fraction (TPF or Pd) values, and Figure 2b shows the false positive fractions (FPF or Pf) values, with mean values shown in bold, surrounded by 5% and 95% quantile curves. Train and test ROC operating points matched up well. There was wide variance for small training sets, but for more than 400 train samples, TPF and FPF values attained asymptotic values.

Figure 3: Generalization error of estimated ROC operating points (shown in Figure 2) as a function of the number of training samples. Generalization errors are differences between the train and test values, $\left(TPF_{train} - TPF_{test}\right)\Big|_{\tau=0}$ and $\left(FPF_{train} - FPF_{test}\right)\Big|_{\tau=0}$. For decision fusion task of fusing two decisions from features with SNR = 1.0, errors reach asymptotic values at approximately 400 samples.

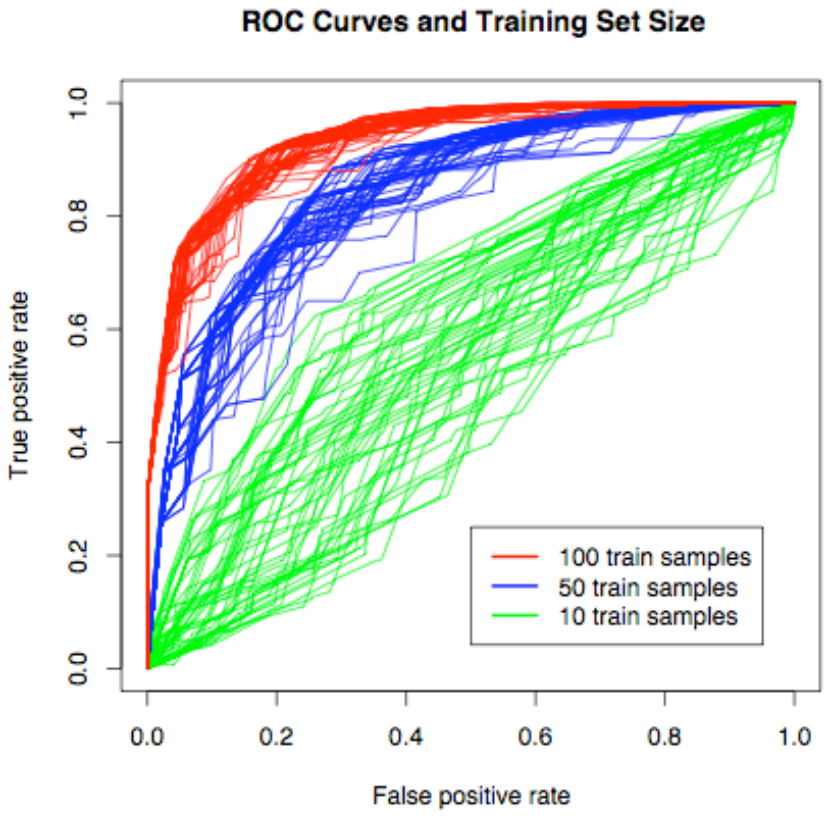


Figure 4: Multiple test ROC curves are shown for fusing five decisions from features with SNR = 2. These ROC curves show testing classification performance; the decision fusion algorithm was trained on training sets of varying sizes and then applied to a constant and independent testing set of 1000 observations. As the number of training cases increased, the test ROC curves rose from a high-variance set around the change diagonal line (10 samples, green curves) to a low-variance set at high performance (100 train samples, red curves). Adequate number of training samples allows the test ROC curves approach their asymptotic shape.

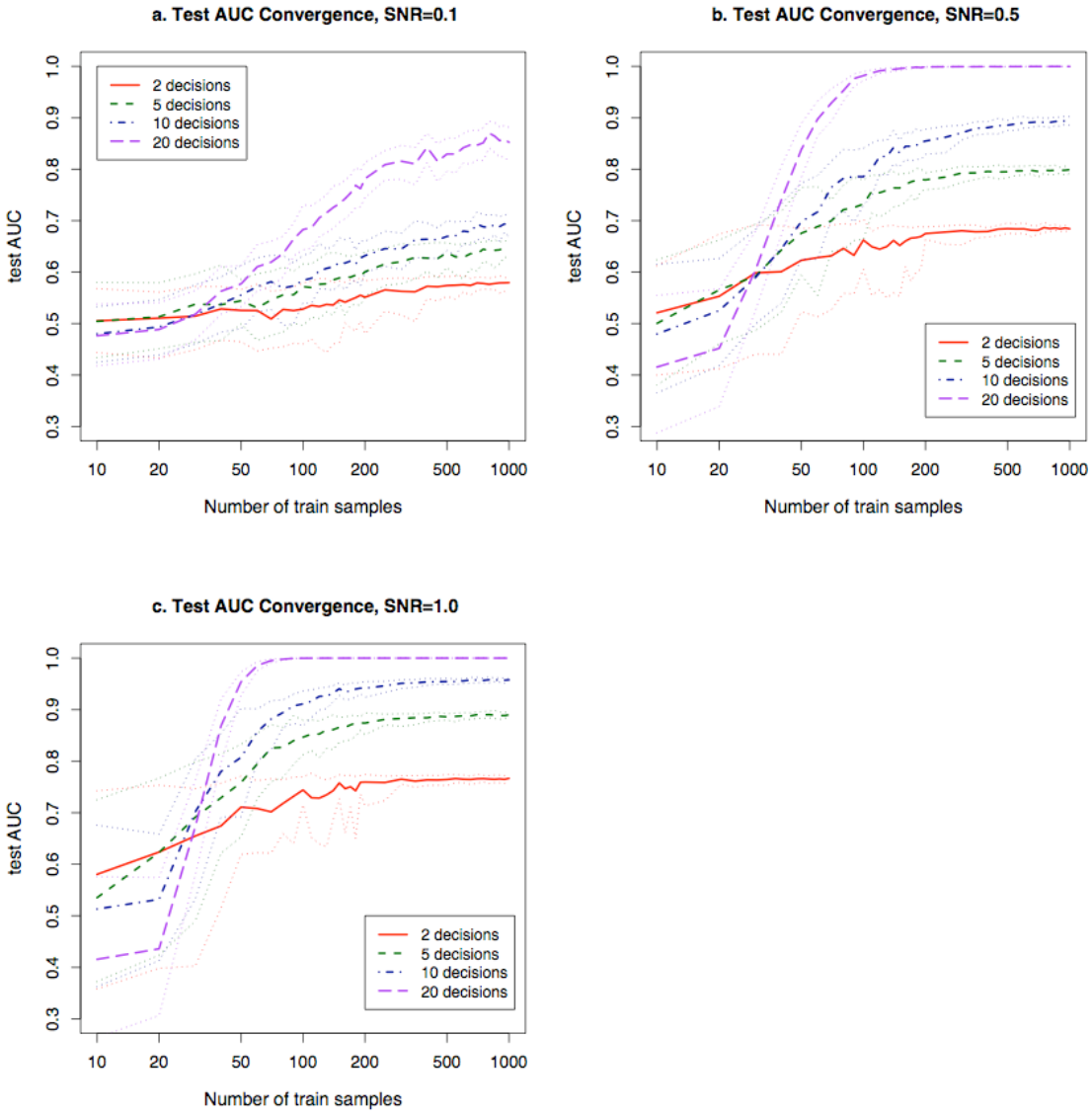


Figure 5: Effect of SNR on testing convergence. Testing AUC versus the number of training samples is plotted for features each with SNR = (a) 0.1, (b) 0.5, and (c) 1.0. For every signal strength scenario, more features provide more information and faster convergence to a higher asymptotic value, since all features have the same SNR value. For example, with 100 features at SNR = 1.0 (Fig. 5c, purple curve), testing AUC levels off with only 50 samples, whereas with the same SNR per feature but fewer features, many hundreds of samples are needed (Fig. 5c, red,

green, and blue curve). As the individual feature SNR increases between the sub-plots, convergence also occurs with fewer training samples.

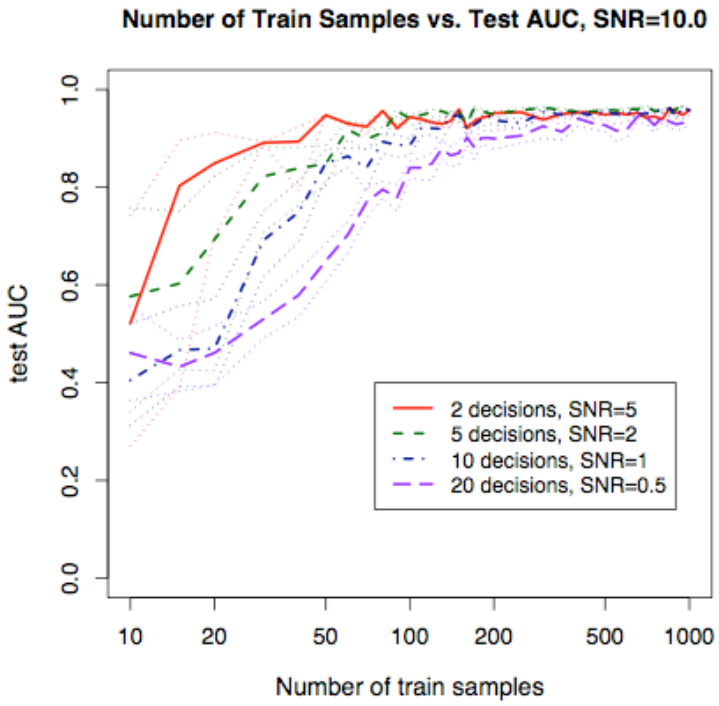


Figure 6: Testing AUC versus the number of training samples, with the total SNR of the whole data set kept constant at SNR = 10. The mean AUC values are plotted in bold, and the 5% and 95% confidence bands are shown in light dotted lines. To measure the effect of diluting the signal across features, the signal was spread evenly across various numbers of features. Having more but weaker features introduced extra uncertainty into the classification problem. Therefore decision fusion required more training samples in order to reach its asymptotic performance value. We used the stopping rule that the asymptote was reached when the AUC was within 0.01 of the asymptote: $\left|AUC_n - AUC_\infty\right| < 0.01$ where AUC_n is the testing AUC for *n* training samples, and AUC_∞ is the asymptotic testing performance. Decision fusion reached the asymptotic testing value $AUC_\infty = 0.95 \pm 0.01$ at 50 training samples for 2 decisions, 90 training samples for 5 decisions, 300 training samples for 10 decisions, and 400 training samples for 20 decisions.

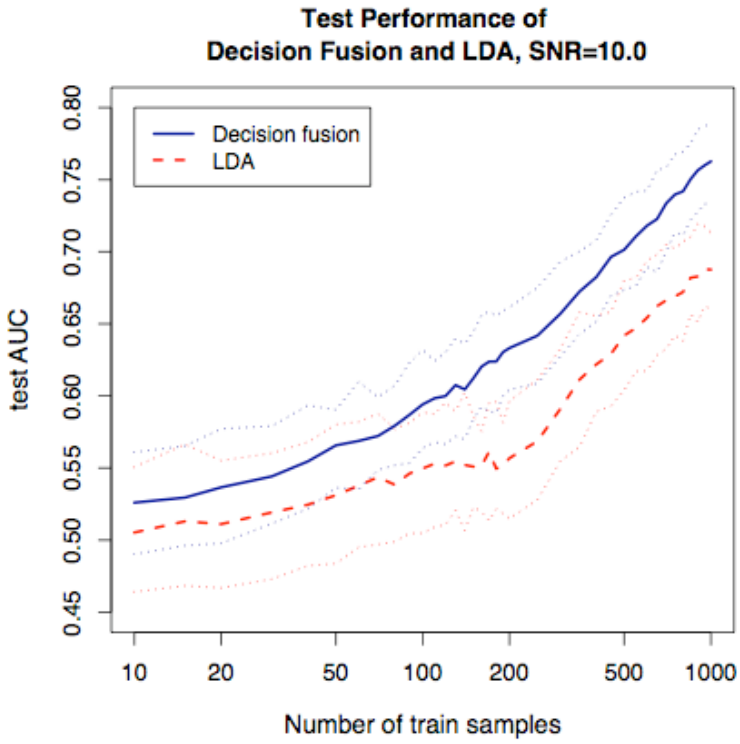
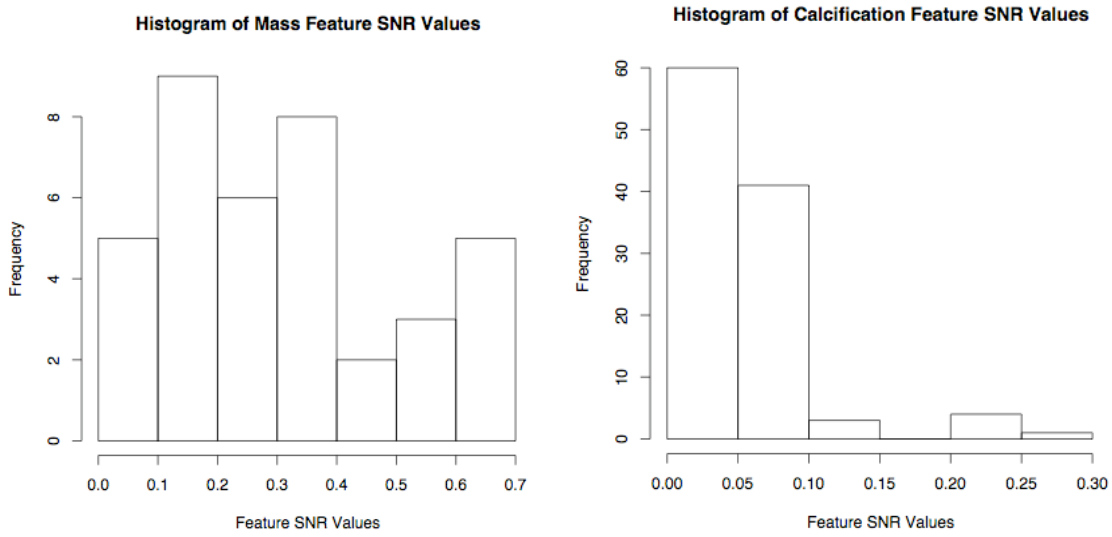


Figure 7: Comparison of classifiers linear discriminant analysis (LDA) and decision fusion using a very large dimensional problem with weak features (total SNR of 10 split among 200 features, each with SNR = 0.05). The classifiers' mean AUC values are shown in bold, with the 5% and 95% percentile bands shown in dotted lines. Decision fusion consistently outperformed LDA on this very weak data set.



(a) SNR histogram for breast mass data set M (b) SNR histogram for breast calcification data set C

Figure 8: Histograms of the feature signal-to-noise ratio (SNR) values for data sets of (a) breast mass lesions and (b) a breast calcification lesions used in our previous study [11]. The calcification data set consisted of much weaker features (lower SNR) and thus presented a much more challenging classification problem. With stronger features, the mass data set presented an easier classification problem.

Bibliography

[1] Fukunaga K, Hayes RR. Effects of sample size on classifier design. IEEE Trans PAMI. 1989;11:873-5.
[2] Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans PAMI. 1991;13:252-64.
[3] Chan H-P, Sahiner B, Wagner RF, Petrick N. Effects of sample size on classifier design for computer-aided diagnosis. Medical Imaging 1998: Image Processing; 1998; San Diego, CA, USA: SPIE; 1998. p. 845-58.
[4] Chan H-P, Sahiner B, Wagner RF, Petrick N, Mossoba JT. Effects of sample size on classifier design: quadratic and neural network classifiers. Medical Imaging 1997: Image Processing; 1997; Newport Beach, CA, USA: SPIE; 1997. p. 1102-13.
[5] Wagner RF, Chan H-P, Sahiner B, Petrick N, Mossoba JT. Finite-sample effects and resampling plans: applications to linear classifiers in computer-aided diagnosis. Medical Imaging 1997: Image Processing; 1997; Newport Beach, CA, USA: SPIE; 1997. p. 467-77.
[6] Sahiner B, Chan H-P, Petrick N, Wagner RF, Hadjiiski LM. Stepwise linear discriminant analysis in computer-aided diagnosis: the effect of finite sample size. Medical Imaging 1999: Image Processing; 1999; San Diego, CA, USA: SPIE; 1999. p. 499-510.
[7] Tourassi GD, Floyd CE. The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis [see comments]. Medical Decision Making. 1997;17(2):186-92.
[8] Chan HP, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers. Medical physics. 1999;26(12):2654-68.
[9] Zheng B, Chang YH, Good WF, Gur D. Adequacy testing of training set sample sizes in the development of a computer-assisted diagnosis scheme. Acad Radiol. 1997;4(7):497-502.
[10] Sahiner B, Chan HP, Petrick N, Wagner RF, Hadjiiski L. Feature selection and classifier performance in computer-aided diagnosis: the effect of finite sample size. Medical physics. 2000;27(7):1509-22.
[11] Jesneck JL, Nolte LW, Baker JA, Floyd CE, Lo JY. Optimized approach to decision fusion of heterogeneous data for breast cancer diagnosis. Medical physics. 2006 Aug;33(8):2945-54.
[12] Tenney RR, Sandell NR. Detection with Distributed Sensors. Aerospace and Electronic Systems, IEEE Transactions on. 1981;AES-17(4):501-10.
[13] Chair Z, Varshney PK. Optimal data fusion in multiple sensor detecton systems. IEEE Transactions on Aerospace and Electronic Systems. 1986;AES-22(1):98.
[14] Reibman AR, Nolte LW. Optimal detection and performance of distributed sensor systems. IEEE Transactions on Aerospace and Electronic Systems. 1987;AES-23(1):24.
[15] Dasarathy BV. Decision fusion strategies in multisensor environments. IEEE Transactions on Systems, Man and Cybernetics. 1991;21(5):1140.

[16] VanTrees HL. Detection, Estimation, and Modulation Theory (Part I). New York: John Wiley & Sons 1968.
[17] Liao Y. Distributed decision fusion in signal detection -- a robust approach. PhD Thesis. 2005.
[18] Bilska-Wolak AO, Floyd CE, Jr, Nolte LW, Lo JY. Application of likelihood ratio to classification of mammographic masses; performance comparison to case-based reasoning. Medical physics. 2003;30(5):949-58.
[19] Bilska-Wolak AO, Floyd CE, Jr., Lo JY, Baker JA. Computer aid for decision to biopsy breast masses on mammography: validation on new cases. Acad Radiol. 2005 Jun;12(6):671-80.
[20] Drakopoulos E, Lee C-C. Optimum multisensor fusion of correlated local decisions. IEEE Transactions on Aerospace and Electronic Systems. 1991;27(4):593.
[21] Vapnik NV. The Nature of Statistical Learning Theory. 2nd Edition ed. New York, NY: Springer-Verlag New York, Inc. 1995.
[22] Metz CE. Basic principles of ROC analysis. Sem Nuc Med. 1978;8:283-98.
[23] Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Statistics in medicine. 1998 May 15;17(9):1033-53.
[24] Metz CE. ROC methodology in radiologic imaging. Investigative Radiology. 1986;21:720-33.
[25] Obuchowski NA, McClish DK. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. Statistics in medicine. 1997 Jul 15;16(13):1529-42.
[26] Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York, NY: Chapman & Hall 1993.

Stacked Generalization in Computer-Assisted Diagnosis Systems: Empirical Comparison of Data Handling Schemes

Georgia D. Tourassi, Jonathan L. Jesneck, Piotr A. Habas, and Maciej Mazurorowski

***Abstract*—Computer-assisted diagnosis (CAD) systems are becoming increasingly popular for the diagnostic interpretation of radiologic images. These CAD systems often involve the stacked generalization of several different decision models. Combining decision models is a common meta-analysis strategy to improve upon the diagnostic performance of each individual model. This study investigates how data handling schemes may affect the performance evaluation of CAD systems that rely on stacked generalization. The study is based on a multistage CAD system for the detection of masses in screening mammograms. The CAD system consists of a series of knowledge-based modules that operate at the Level 0 capturing morphological as well as multiscale textural information. Then, the knowledge-based predictions are combined with a Level 1 classifier. The study shows that a leave-one-out sampling scheme appears to be an effective and relatively unbiased strategy to estimate the overall performance of a CAD system that is based on stacked generalization. However, extra caution should be placed on the complexity of the Level 1 combiner. When the available dataset is relatively small, a relatively simple learning system such as a backpropagation neural network with very few hidden nodes is preferable to avoid optimistically biased estimates of diagnostic performance.**

I. Introduction

STACKED generalization is a popular technique to combine multiple decision models in an attempt to improve classification performance [1]. The individual decision models are considered the first level of analysis. The outputs of these models are then combined with a level one generalizer to improve upon the performance of the individual models. Several studies have confirmed the effectiveness of stacked generalization, although the amount of improvement is strongly dependent on the actual database and problem at hand [2].

Manuscript received January 31, 2007. This work was supported in part by the National Cancer Institute under Grant R01 CA101911.

Georgia D. Tourassi Ph.D. is with the Duke Advanced Imaging Laboratories, Department of Radiology, Duke University Medical Center, Durham, NC 27705 (phone: 919-684-1447; fax: 919-684-1491; e-mail: georgia.tourassi@duke.edu).

Jonathan L. Jesneck, B.S.E.E is with Duke Advanced Imaging Laboratories, Department of Radiology and the Department of Biomedical Engineering, Duke University, Durham, NC 27705 (e-mail: jonathan.jesneck@duke.edu).

Piotr A. Habas, M.S.E. is with the Computational Intelligence Laboratory, Department of Electrical and Computer Engineering, University of Louisville, Louisville KY 40292 (e-mail: piotr.habas@louisville.edu).

Maciej Mazurorowski, B.S.E.E. is with the Computational Intelligence Laboratory, Department of Electrical and Computer Engineering, University of Louisville, Louisville KY 40292 (e-mail: maciej.mazurorowski @louisville.edu).

In particular, stacked generalization is applied extensively in medical computer-assisted decision (CAD) systems for the diagnostic interpretation of radiologic images. CAD systems typically follow a modular approach. The different modules are designed to operate on different subsets of features (e.g., morphological, textural, clinical, etc.) and/or employ different types of decision algorithms such neural networks, support vector machines, decision trees etc. These CAD modules are considered the first level classifiers. Then, a higher-level decision model is used to combine the lower-level outputs to achieve greater diagnostic performance [e.g., 3-10].

There are several choices for combining the lower level decisions. These choices range from simple voting schemes such as average, maximum, and majority vote to more elaborate learning systems. Specifically, artificial neural networks are a popular choice for high level decisions because they can capture and model complex relationships among the lower level decision models.

Although there are numerous published studies on stacked generalization, most of these studies focus on two key issues: (i) how to choose the lower level models that serve as its inputs, and (ii) how to choose the higher-level model. However, as Wolpert stated in his defining paper [1], there are no set rules regarding these decisions and stacked generalization remains to a large extent a "black-box" operation. Consequently, the effectiveness of combining classifiers strongly depends on the particular problem at hand.

The purpose of this study is to address the issue of data handling when building modular CAD systems that rely on stacked generalization. In general, data handling is a critical issue in CAD since limited availability of clinical data is a common restriction. Therefore, CAD researchers rely on sophisticated data handling schemes such as leave-one-out crossvalidation or bootstrapping to capitalize on the available data [11,12]. Previous studies suggest that these handling schemes may lead to optimistic estimates of predictive performance when limited datasets are reused not only for train/testing complex decision models but also for optimizing the feature selection process as well as the architecture and training parameters of the decision models [13,14]. The data handling issues becomes increasingly more important with multilevel CAD systems. The same available data need to be used effectively to train and test the full CAD system without introducing any optimistic bias.

With multiple classifiers stacked at different levels, the risk of introducing a positive bias by reusing data across levels is real when evaluating the diagnostic performance of the full CAD system. We aim to address this concern with respect to our own CAD system for the detection of masses in screening mammograms. This is an empirical investigation of how the data-handling scheme may affect the reported performance of a CAD system that relies on stacked generalization to improve its diagnostic performance.

II. The CAD System

A. Overview

Previously we presented a knowledge-based CAD (KB-CAD) system for the detection of masses in screening mammograms. The details of this system are provided elsewhere [15,16]. This is a brief description.

The knowledge-based CAD system is designed to provide a second opinion regarding the presence or absence of a mass at a specific mammographic location that is under scrutimization. A 512x512 pixel region of interest (ROI) is extracted around the indicated location. The query ROI is compared to a knowledge database of mass and normal ROI templates with known ground truth. A decision is made based on how well the query ROI matches the mass templates relative to those templates depicting normal breast parenchyma. Template matching is based on information-theoretic principles such as mutual information [17]. The underlying hypothesis of this detection scheme is that if the query region contains a mass, it should match the mass templates better than the normal templates, thus resulting in a higher decision index.

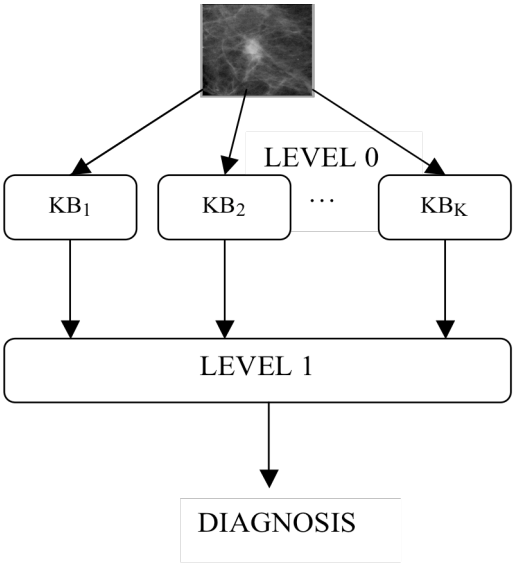


Fig. 1. Schematic representation of a CAD system composed of stacking several knowledge-based decision algorithms.

Although the mutual information between two ROIs could be measured directly without any image preprocessing, in a previous study we reported that entropy-based filtering of the ROIs could substantially contribute to improving the diagnostic performance of the system [18]. Therefore, we proposed a two-level CAD system by stacking together several KB-CAD modules (Fig. 1).

B. Level 0 Classifiers

The Level 0 classifiers are essentially seven different KB-CAD systems operating on the same query ROI that is preprocessed differently. The first KB-CAD system operates on the unprocessed ROI while the remaining operate on different preprocessed versions of the same ROI. Specifically, the query ROI is preprocessed with an entropy-based filter. The filter essentially replaces the intensity value of each ROI pixel with a new value that captures the local image entropy around the pixel. The filtering step is repeated at several scales by varying the neighborhood size of the entropy-based filter (3x3, 7x7, 9x9, 11x11, 15x15, and 21x21 pixels). Therefore, the Level 0 classifiers are designed to capture intensity-based similarity as well as multiscale textural similarity between the query and the templates stored in the knowledge database. For each quer ROI and its preprocessed versions, a separate prediction is made using the series of KB-CAD systems.

C. Level 1 Classifier

The next step is to combine the Level 0 predictions into one final prediction regarding the query. Since the outputs of the Level 0 decision models is a continuous value rather than a binary decision, there are several options for meta-analysis. Some commonly used choices are the maximum value, minimum value, the average value, or some weighted average of the $KB\text{-}CAD_k$ predictions.

Backpropagation neural networks (BP-ANN) are a popular choice for combining predictions due to their ability to capture complex, non-linear relationships among the various decision models [19,20]. A BP-ANN was constructed to combine the seven predictions into one comprehensive decision regarding the presence or absence of mass. The neural network had a three-layer, feed-forward architecture. Experiments with variable number of hidden nodes were performed to assess whether the conclusions of the study are affected by the complexity of the BP-ANN. The BP-ANNs were constructed using the JMP Software (available from SAS, Cary, NC). In addition, the maximum, minimum, and average prediction decision models were applied for comparison.

III. Database

The available database for this study consisted of screen-film mammograms selected from the Digital Database for Screening Mammography (DDSM) [21]. Cases from the Lumisys volumes were selected and 512x512 pixel ROIs

were automatically extracted around annotated malignant and benign masses. In addition, normal ROIs were extracted from normal mammograms and from the cance-free breasts of abnormal mammograms that did not contain any annotations. In total, there were 1,820 ROIs available. Of those, there were 901 ROIs depicting a mass and 919 ROIs depicting normal breast parenchyma.

IV. Experimental Design

The purpose of a typical CAD study is to capitalize on the available data so that the system can be effectively trained and tested without compromising its ability to generalize. For the specific CAD system at hand, data is required for the following tasks: 1) build the knowledge database for the Level 0 KB-CAD modules, 2) test the KB-CAD modules, 3) train the Level 1 neural network, and 4) test the neural network and ultimately the full multi-level CAD system. Note that with the simpler maximum, minimum, and average vote decision models, task 3 is obsolete since no training is necessary. Ideally, different subsets of ROIs should be used to achieve the above tasks without introducing any optimistic biases.

Since the available database is limited, it is difficult to know the optimal way to use the available data to effectively achieve the above four tasks. Using as many ROIs as possible to build the knowledge database for the Level 0 CAD modules is critical for accurate generalization to new ROIs. Building a diverse and comprehensive knowledge database is a critical component of an effective CAD system [22]. However, reserving too much of the available data for building the Level 0 classifiers would reduce the ability of the Level 1 BP-ANN to train effectively and learn important relationships among the seven KB-CAD modules. Therefore, the generalization performance of the ANN could be severely compromised.

We experimented with the implementation of various data handling schemes to assess the impact of data handling on the reported results. Starting with the original database of 1,820 ROIs, the database was randomly split in 3 sets of roughly equal size and equal mass prevalence. Table I provides the relevant statistics.

TABLE 1: Dataset Statistics

DATASET	MAIGNANT MASSES	BENIGN MASSES	NORMAL
SET 1	166	137	304
SET 2	156	139	312
SET 3	167	136	303
ALL	489	412	919

Then, the following data handling schemes were implemented:

SCHEME 1: Set 1 was used as the knowledge database of the Level 0 KB-CAD modules. These modules were then tested on Set 2. The $KB\text{-}CAD_k$ outputs on Set 2 were used as the training data for the Level 1 ANN. Finally, the two-level CAD system was tested on Set 3 which was reserved from the beginning as the validation set. Scheme 1 is essentially the preferred strategy if there are enough available data since it keeps the training/testing of each level independent from one another. The obvious drawback of this scheme is that only 1/3 of the database is available for each task.

SCHEME 2: Sets 1 and 2 were used for developing and testing the level 0 KB-CAD systems based on the leave-one-out sampling scheme [12]. In other words, each ROI from the 1,214 ROIs in sets 1 and 2 was excluded once to serve as the query while the remaining 1,213 served as the knowledge database. The process was repeated until each ROI served as the query. The 1,214 KB-CAD outputs were used as the training set for the Level 1 BP-ANN. The trained BP-ANN was finally tested on set 3 which was reserved from the beginning as the validation set. The advantage of this scheme is that it maximizes on the available cases for training/testing the Level 0 classifiers and for training the Level 0 BP-ANN. The drawback is that the final validation set is still limited to 1/3 of the available data (i.e., set 3).

SCHEME 3: The final scheme employs the leave-one-out sampling scheme for both the level 0 and level 1 classifiers. Although this scheme capitalizes on the available data, it is unclear whether it introduces any optimistic bias by essentially reusing the available cases in a cascading format.

V. Results

A. Performance of Level 0 KB-CAD Modules

The diagnostic performance of all classifiers was assessed using Receiver Operating Characteristics (ROC) analysis [23]. The reported index of performance is the area under the ROC curve. Generally, a higher area index reflects a better diagnostic performance. The ROC area index is considered a more appropriate performance criterion for medical diagnostic problems because it is independent of the decision threshold and the disease prevalence. We used the ROCKIT algorithm developed by Metz *et al.* to fit and compare ROC curves [24,25].

Table II summarizes the performance of the Level 0 classifiers depending on the data-handling scheme. The table shows the same general trend. As the size of the knowledge database increases (from scheme 1 to scheme 3), so does the performance of the KB-CAD modules. This is particularly true for some modules (e.g., $KB\text{-}CAD_3$, $KB\text{-}CAD_4$, $KB\text{-}CAD_5$, $KB\text{-}CAD_6$) where statistically significant improvement was observed with a two-tailed p-value < 0.05.

TABLE II: Testing ROC area index for all level 0 classifiers depending on the data-handling scheme

MODULE	SCHEME 1	SCHEME 2	SCHEME 3
KB-CAD$_1$	0.83±0.02	0.87±0.02	0.87±0.01
KB-CAD$_2$	0.67±0.02	0.68±0.02	0.69±0.01
KB-CAD$_3$	0.78±0.02	0.78±0.02	0.81±0.01
KB-CAD$_4$	0.78±0.02	0.78±0.02	0.82±0.01
KB-CAD$_5$	0.76±0.02	0.76±0.02	0.80±0.01
KB-CAD$_6$	0.72±0.02	0.75±0.02	0.75±0.01
KB-CAD$_7$	0.76±0.02	0.78±0.02	0.79±0.01

B. Performance of Level 1 Combiner

Similarly, the diagnostic performance of the BP-ANN was also assessed using ROC analysis based on the three data handling schemes. The BP-ANN performance is shown in Table III. Results are reported for three different sizes of the hidden layer (i.e., 3,6,9 hidden nodes). BP-ANN$_j$ denotes a network with j hidden nodes. Note that the training parameters of the BP-ANNs were kept fixed across all data handling schemes. Specifically, all networks were trained with 50 training iterations and 0.001 overfit penalty.

In addition, the Table includes the ROC performance of the simple Level 1 combiners (i.e., maximum, minimum, and average prediction). The simpler combiners generate a decision index DI by taking the maximum (Eq. 1), minimum (Eq. 2), or average (Eq. 3) value of the seven KB-CAD predictions on the dataset designated for final validation depending on the data-handling scheme.

$$DI_{\max} = \max\{KB\_CAD_j\} \; where \; j = 1,2,...,7 \quad [1]$$

$$DI_{\min} = \min\{KB\_CAD_j\} \; where \; j = 1,2,...,7 \quad [2]$$

$$DI_{avg} = \frac{\sum_{j=1}^{7} KB\_CAD_j}{7} \quad [3]$$

TABLE III: Testing ROC area index for the Level 1 classifier depending on the data-handling scheme

LEVEL 1	SCHEME 1	SCHEME 2	SCHEME 3
BP-ANN$_3$	0.89±0.02	0.92±0.01	0.92±0.01
BP-ANN$_6$	0.86±0.02	0.91±0.01	0.93±0.01
BP-ANN$_9$	0.84±0.02	0.89±0.01	0.93±0.01
Maximum	0.87±0.02	0.89±0.02	0.87±0.01
Minimum	0.76±0.02	0.76±0.02	0.78±0.01
Average	0.88±0.02	0.90±0.02	0.88±0.01

Figure 2 shows the corresponding ROC curves of the CAD system under all three data handling schemes and for all Level 1 combiners. Although the reported ROC performance for Scheme 1 and Scheme 2 is based on the validation set 3, the ROC performance for Scheme 3 is based on the full dataset, as previously explained.

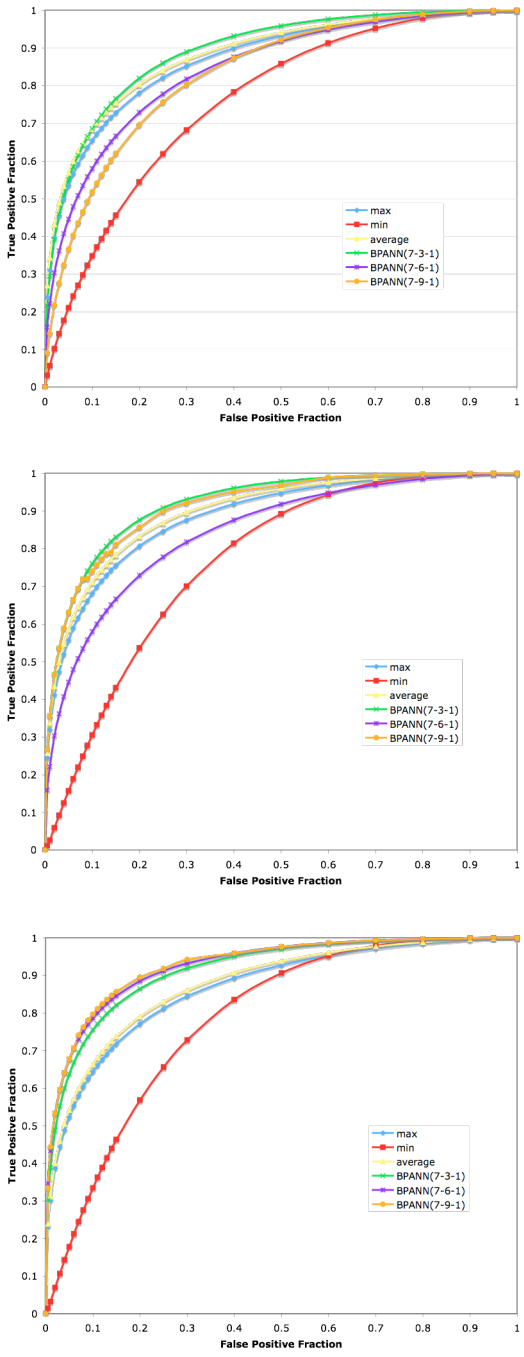


Figure 2: ROC performance of Level 2 combiners accrding to the three data handling scheme investigated: (a) scheme 1, (b) scheme 2, (c) scheme 3.

Overall, the performance of the BP-ANN improved as the size of the training set increased from 607 (for Scheme 1), to 1,214 (for Scheme 2), to 1,819 (for Scheme 3). The increase was statistically significant, particularly for the more

complex BP-ANNs. This is not surprising considering that as the number of hidden nodes increases, more training examples are necessary to improve the genaralizabilty of the BP-ANN. In contrast, the simple combiners had inferior performance compared to the BP-ANN under all data handling schemes. This result confirms that non-linear networks are more effective as Level 1 stacking combiners.

VI. Conclusions

We have addressed the issue of data handling when using stacked generalization to improve the diagnostic performance of CAD systems. Our empirical study was based on a typical multistage CAD system developed for the detection of masses in screening mammograms. The CAD system relied on knowledge-based modules that operate at the Level 0 capturing morphological as well as multiscale textural information. Then, the knowledge-based predictions were combined with a Level 1 classifier. Our study showed the following.

First, BP-ANNs are more effective Level 1 combiners than the simpler choices such as maximum, minimum, and average combiners. Although simple, these combiners do not make use of any Level 1 learning. In contrast, BP-ANNs capitalize on subtle, yet important relationships that exist among the Level 0 predictions. Second, the complexity of the BP-ANNs should be dictated by the amount of available data. Simpler BP-ANN architectures should be employed for limited datasets to avoid issues of under-training and poor generalization. Finally, a leave-one-out sampling scheme appears to be an effective and relatively unbiased strategy to estimate the overall performance of a CAD system that is based on stacked generalization. However, extra caution should be placed on the complexity of the Level 1 combiner. When the available dataset is relatively small, a relatively simple classifier such as a BP-ANN with very few hidden nodes is preferable to avoid optimistically biased estimates of diagnostic performance.

VII. Acknowledgements

We would like to thank Brian Harrawood, B.S. for scientific programming.

References

[1] Wolpert, D.H., Stacked Generalization, Neural Networks, Vol. 5, pp. 241-259, 1992.

[2] Ting, K.M. and Witten, I.H., Issues in Stacked Generalization, *Journal of Artificial Intelligence Research* Vol. 10, pp. 271-289, 1999.

[3] Sehgal, M.S.B. Gondal, I. Dooley, L., Support vector machine and generalized regression neural network based classification fusion models for cancer diagnosis, Hybrid Intelligent Systems, 2004. HIS '04. Fourth International Conference, pp. 49- 54, 2004.

[4] Sun, Y. Babbs, C. Delp, E.J., Full-field mammogram analysis based on the identification of normal regions, IEEE International Symposium on Biomedical Imaging: Macro to Nano 2004, Vol. 2, pp. 1131- 1134, 2004.

[5] Bovis K., Singh S., Classification of Mammographic Breast Density Using a Combined Classifier Paradigm, International Workshop on Digital Mammography, 2002.

[6] Campanini R., Dongiovanni D., Iampieri E., et a., A novel featureless approach to mass detection in digital mammograms based on support vector machines, *Phys. Med. Biol.* 49 961-975, 2004.

[7] Wei L., Yang Y., Nishikawa R.M., Jiang Y., A study on several Machine-learning methods for classification of Malignant and benign clustered microcalcifications, IEEE Trans Med Imag Vol. 24(3), pp. 371- 380, 2005.

[8] Lo J.Y., Gavrielides M., Markey M.K., Jesneck J.L., Computer-aided classification of breast microcalcification clusters: Merging of features from image processing and radiologists, SPIE Medical Imaging Conference, 2003.

[9] Tourassi G.D., Frederick E.D., Floyd C.E. Jr, Coleman R.E., Multifractal texture analysis of perfusion lung scans as a potential diagnostic tool for acute pulmonary embolism, Comput Med Biol, Vol. 31(1), pp. 15-25, 2001.

[10] Suzuki K., Armato S.G., Li F., Sone S., Doi K., Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography, Med Phys, Vol. 30(7), pp. 1602-1617, 2003.

[11] Efron B., Tibshirani R.J., *An Introduction to the Bootstrap*, *Monographs on Statistics and Applied Probability*, ed. D. R. Cox et al. (Chapman & Hall, New York, NY, 1993).

[12] Efron B., Gong G., A leisurly look at the bootstrap, the jackknife, and cross-validation, The American Statistician, Vol. 37(1), pp. 36-48, 1983.

[13] Chan H.P. Sahiner G., Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers, Med Phys, Vol. 26(12), pp. 2654-2668, 1999.

[14] Sahiner B., Chan H.P., Petrick N., Wagner R.F., Hadjiiski L., Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size, Med Phys, Vol. 27(7), pp. 1509-1522, 2000.

[15] Tourassi G.D., Vargas-Voracek R., Catarious D.M., Floyd C.E., Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information, Med Phys, Vol. 30 pp. 2123-2130, 2003.

[16] Tourassi G.D., Harrawood B., Singh S. Lo J.Y., Floyd C.E., Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammogram, Med Phys, Vol. 34(1), pp. 140-150, 2007.

[17] Cover T.M., Thomas J.A., *Elements of Information Theory*, John Wiley & Sons, New York, 1991.

[18] Tourassi G.D., Bilska-Wolak A., Habas P.A., Floyd C.E., Incorporation of a Multi-Scale Texture-Based Approach to Mutual Information Matching for Improved Knowledge-Based Detection of Masses is Screening Mammograms, to be presented at the 2007 SPIE Conference on Medical Imaging.

[19] Rumelhart D.E., Hinton G.E., Williams R.J., Learning internal representations by error propagation, In: Rumelhart DE, McClelland JL, ed. *Parallel Distributed Processing: Explorations in the Microstructures of Cognition.* Cambridge, MA: The MIT Press, vol. I, pp. 318-362, 1986.

[20] Simpson P.K., Artificial Neural Systems: foundations, paradigms, applications, and implementations. Neural Networks: Research and Applications, ed. S. Amari, et al. Elmsford, NY: Pergamon Press, 1990.

[21] Heath M., Bowyer K., Kopans D., et al, Current Status of the Digital Database for Screening Mammography, *in Digital Mammography*, Kluwer Academic Publishers, 1998.

[22] Tourassi G.D., Floyd C.E. Jr., Knowledge-Based Detection of Mammographic Masses: Analysis of the Impact of Database Comprehensiveness, Proc SPIE, Vol. 5748, pp. 399-405, 2005.

[23] Obuchowski N.A., Receiver Operating Characteristic curves and their use in radiology, Radiology, Vol. 229, pp. 3-8, 2003.

[24] Metz C.E., Shen J.-H., Herman B.A., New methods for estimating a binormal ROC curve from continuously-distributed test results, presented at the 1990 joint meetings of the American Statistical Society and the Biometric Society. Anaheim, CA, August 1990.

[25] Rockette H.E., Gur D., Metz C.E., The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques, Investigative Radiology, Vol. 27, pp. 169-172, 1992.

BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–7

Identifying circulating protein markers for breast cancer detection in premenopausal women

Jonathan L. Jesneck [1−4], Sayan Mukherjee [2,5,6], Anna E. Lokshin [7,8], Jeffrey R. Marks [5,9], Merlise Clyde [2], Joseph Y. Lo [1,4,10*]

[1]Department of Biomedical Engineering, [2]Institute of Statistics and Decision Sciences, [3]Computational Biology and Bioinformatics Program, [4]Duke Advanced Imaging Labs, Department of Radiology [5]Institute for Genome Sciences and Policy, [6]Department of Computer Science, Duke University [7]Division of Hematology/Oncology, [8]University of PIttsburgh School of Medicine and University of Pittsburgh Cancer Institute, Pittsburgh, PA 15213, [9]Department of Experimental Surgery, Duke University Medical Center, [10]Medical Physics Graduate Program, Duke University, Durham NC 27708, USA

ABSTRACT

Motivation: Screening mammography for breast cancer is less effective for younger women. To complement mammographic screening for premenopausal women, we investigated the feasibility of a diagnostic blood test using serum proteins. This study used a set of 98 serum proteins and choose diagnostically relevant subsets via various feature-selection techniques. To account for model selection uncertainty, we applied iterated Bayesian model averaging because of its good generalization performance and tendency to select a small set of features. We assessed generalization performance using leave-one-out cross-validation (LOOCV) and receiver operating characteristic (ROC) curve analysis.

Results: The classifiers were able to distinguish normal tissue from breast cancer with a classification performance of $AUC = 0.80$ with the proteins MIF, MMP-9, and MPO. The classifiers separated normal tissue from benign lesions similarly at $AUC = 0.78$. However, the serum proteins of benign and malignant lesions were indistinguishable ($AUC = 0.55$). The classification tasks of normal vs. cancer and normal vs. benign selected the same top feature: MIF, which indicates inflammatory response. Overall, the considered features showed promise in detecting lesions but are probably more indicative of secondary effects rather than specific for malignancy.

Availability: The software and data set are available at http://deckard.duhs.duke.edu/resources.html.

Contact: joseph.lo@duke.edu

1 INTRODUCTION

Breast cancer is unfortunately a very common and lethal disease. It strikes one in eight women (Amer. Cancer Soc., 2007), accounts for one-third of all cancer diagnoses (Lacey *et al*., 2001), and is the second leading cause of cancer death among American women (Ferrini *et al*., 1996). For American women in 2007, Jemal *et al*. (2007) estimate 178,480 new breast cancer cases and 40,460 deaths.

Despite such discouraging morbidity and mortality statistics, breast cancer patients significantly improve their chance of survival with early diagnosis and treatment (Cady and Michaelson, 2000). Currently the preferred screening tool is mammography. However, screening mammography suffers from only moderate sensitivity rates (estimated at 75% to 90%) (Ferrini *et al*., 1996) and high false positive rates, with only 13-29% of suspicious masses determined to be malignant by biopsy (Meyer *et al*., 1984; Rosenberg *et al*., 1987; Yankaskas *et al*., 1988). While mammography's positive predictive value (PPV) ranges from 60% to 80% in older women (age 50-69), it is only 20% in women under age 50 (Ferrini *et al*., 1996).

Mammographic screening is more problematic for younger women (Tabar *et al*., 1995; Kerlikowske *et al*., 1995), whose denser breast tissue occludes lesions. Premenopausal women account for approximately one third of breast cancer patients in Britain and other high-risk countries (Simpson *et al*., 1988). Younger patients tend to experience more aggressive forms of breast cancer and have significantly lower survival rates and higher local and distant relapse rates than older patients (de la Rochfordiere and Asselain, 1993). In fact, Dubsky *et al*. (2002) found that, after lymph node status, young age was the second most powerful risk factor for breast cancer recurrence and mortality.

To boost the diagnostic performance in younger women, mammographic screening can benefit from additional and complementary technologies, such as protein profiling and gene expression profiling. Although some progress has been made using gene expression profiling of excised breast tissue samples (van't Veer *et al*., 2001; Perou *et al*., 1999; Gruvberger *et al*., 2001; Martin *et al*., 2000; Zajchowski *et al*., 2001; Sorlie *et al*., 2001; West *et al*., 2001), collecting such data requires invasive biopsies, which is less practical for screening programs, and it is unclear what advantage these invasive tests would have over routine histopathological analysis of biopsy samples.

For the far less invasive blood draw, however, it is fitting and convenient to perform protein profiling. Proteins offer detailed information about tissue health conditions, allowing the identification of cancer type and risk, and thereby prompting potentially better targeted and more effective treatment. Serum and plasma protein-based screening tests have already been developed for many diseases, such as Alzheimer's disease (Hye *et al*., 2006), cardiovascular disease (Wang *et al*., 2006), prostate cancer (Polascik *et al*., 1999), and ovarian cancer (Gorelik *et al*., 2005).

*To whom correspondence should be addressed

Table 1. Subject demographics.

	Normal	Benign	Malignant	Total
Number of subjects	68 (41%)	48 (29%)	49 (30%)	165 (100%)
Mean age (years)	36 ± 8	38 ± 9	42 ± 4	38 ± 8
Race: Black	23 (41%)	19 (34%)	14 (25%)	56 (34%)
Race: White	45 (41%)	29 (27%)	35 (32%)	109 (66%)

Table 2. List of the 98 serum proteins measured by ELISA assay (Luminex platform).

ACTH	FSH	IP-10	PROLACTIN
Adiponectin	G-CSF	Kallikrein 10	RANTES
AFP	GH	Kallikrein 8	Resistin
Angiostatin	GM-CSF	Leptin	S-100
Apolipoprotein A1	GROa	LH	SAA
Apolipoprotein Apo A2	Haptoglobin	MCP-1	SCC
Apolipoprotein Apo B	HGF	MCP-2	sE-Selectin
Apolipoprotein Apo C2	IFN-a	MCP-3	sFas
Apolipoprotein Apo C3	IFN-g	Mesothelin(IgY)	sFasL
Apolipoprotein Apo E	IGFBP-1	MICA	sICAM-1
CA 15-3	IL-10	MIF	sIL-6R
CA 19-9	IL-12p40	MIG	sVCAM-1
CA-125	IL-13	MIP-1a	TGFa
CA72-4	IL-15	MIP-1b	TNF-a
CD40L (TRAP)	IL-17	MMP-1	TNF-RI
CEA	IL-1a	MMP-12	TNF-RII
Cytokeratin 19	IL-1b	MMP-13	tPAI-1
DR5	IL-1Ra	MMP-2	TSH
EGF	IL-2	MMP-3	TTR
EGFR	IL-2R	MMP-7	ULBP-1
EOTAXIN	IL-4	MMP-8	ULBP-2
ErbB2	IL-5	MMP-9	ULBP-3
FGF-b	IL-6	MPO	VEGF
Fibrinogen	IL-7	NGF	
Fractalkine	IL-8	PAI-I(active)	

For breast cancer, however, there are currently very few markers used clinically. Limited success has been reported for identifying breast cancer proteins using mass spectroscopy of the tumor tissue (Kreunin *et al*., 2007), but to date no strong protein biomarkers for breast cancer have been found in serum. Some studies have shown correlations between individual circulating proteins and breast cancer (Malkas *et al*., 2006), but to our knowledge these promising proteins have not been assessed collectively to identify the most promising subset for breast cancer detection.

The goals of this study were to identify promising serum proteins to detect breast cancer and to investigate the feasibility of using these protein levels in a screening tool based on statistical models. For improved predictive performance on this noisy data set, we used iterated Bayesian model averaging (Hoeting *et al*., 1999) of classical regression models (linear, logistic, and probit). To better understand the cancer-specificity of the screening test, we also ran the classifiers on proteins of benign lesions and of normal breast tissue.

2 METHODS

2.1 Data Collection

This study enrolled 165 women undergoing diagnostic biopsy at Duke University Medical Center for breast cancer from June 1999 to October 2005. Table 1 shows the demographics of the study population.

Blood sera were collected under the HIPAA-compliant protocol "Blood and Tissue Bank for the Discovery and Validation of Circulating Breast Cancer Markers." Blood was collected from subjects prior to surgical resection. All specimens were collected in red-stoppered tubes and processed generally within 4 hours (but not greater than 12 hours) after collection and stored at $-80°$C. Sera were assayed using the Luminex platform and reagents for the 98 proteins shown in Table 2. In addition to the protein levels, patient age and race were also recorded.

2.2 Regression with Variable Selection

In order to incorporate these proteins into a breast cancer screening tool, we must build statistical models linking the protein levels to the probability of malignancy. We used the following three common regression models: linear regression $Y_i = X_i\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2)$, logistic regression $\Pr(Y_i = 1|\beta) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$, and probit regression $\Pr(Y_i = 1|\beta) = \Phi(X_i\beta)$, where Y is the response vector, X is the matrix of observed data, β is the vector of coefficients, ϵ is additive noise, and $\Phi(\cdot)$ is the cumulative distribution function of the normal distribution.

These models become unstable and predict poorly if there are relatively few observations and many features. It is better to choose a subset useful features, but when the number of features, p, is large, it is computationally infeasible to compare the full set of 2^p models. Various feature-selection techniques navigate the model search space, either by a deterministic heuristic, such as stepwise feature selection, or stochastically, such as Bayesian model selection (Lee *et al*., 2003; Sha *et al*., 2004).

Regardless of the variable selection method, choosing only one model for prediction comes with an inherent risk. When multiple possible statistical models fit the observed data similarly well, it is risky to make inferences and predictions based only on a single model (Hoeting *et al*., 1999). In this case predictive performance suffers, because standard statistical inference typically ignores model uncertainty.

2.3 Accounting for Model Uncertainty

We accounted for model-selection ambiguity by using a Bayesian approach to average over the possible models. We considered a set of models $M_1, \ldots, M_B$, where each model M_k consists of a family of distributions $\{\mathrm{p}(D|\theta_k, M_k)\}$ indexed by the parameter vector θ_k, where $D = (X, Y)$ is the observed data. Using a Bayesian method (Hodges, 1987; Draper, 1995; Hoeting *et al*., 1999; Berger and Pericchi, 2001; Chipman *et al*., 2001; Clyde and George, 2004) to average over the set of considered models, we first assigned a prior probability distribution $\mathrm{p}(\theta_k|M_k)$ to the parameters of each model M_k. This formulation allows a conditional factorization of the joint distribution,

$$\mathrm{p}(D, \theta_k, M_k) = \mathrm{p}(D|\theta_k, M_k)\ \mathrm{p}(\theta_k|, M_k)\ \mathrm{p}(M_k). \quad (1)$$

Splitting the joint distribution in this way allowed us to implicitly embed the various models inside one large hierarchical mixture model. This form allowed us to fit these models using the computational machinery of Bayesian model averaging.

2.4 Bayesian Model Averaging

Bayesian model averaging (BMA) accounts for model uncertainty by averaging over the posterior distributions of multiple models, allowing for more robust predictive performance. If we are interested in predicting a future observation D_f from the same process that generated the observed data D, then we can represent the predictive posterior distribution $\mathrm{p}(D_f|D)$ as an average of over the models, weighted by their posterior probabilities

(Rafftery, 1995; Rafftery *et al*., 1995; Hoeting *et al*., 1999):

$$\mathrm{p}(D_f|D) = \sum_k \mathrm{p}(D_f|D, M_k)\ \mathrm{p}(M_k|D) \quad (2)$$

where the sum's first term $\mathrm{p}(D_f|D, M_k)$ is a posterior weighted mixture of conditional predictive distributions

$$\mathrm{p}(D_f|D, M_k) = \int \mathrm{p}(D_f|\theta_k, M_k)\ \mathrm{p}(\theta_k|D, M_k)d\theta_k, \quad (3)$$

and the sum's second term $\mathrm{p}(M_k|D)$ is a model's posterior distribution

$$\mathrm{p}(M_k|D) = \frac{\mathrm{p}(D|M_k)\ \mathrm{p}(M_k)}{\sum_k \mathrm{p}(D|M_k)\ \mathrm{p}(M_k)}, \quad (4)$$

which incorporates the model's marginal likelihood

$$\mathrm{p}(D|M_k) = \int \mathrm{p}(D|\theta_k, M_k)\ \mathrm{p}(\theta_k|M_k)d\theta_k. \quad (5)$$

2.5 Promoting Computational Efficiency by Considering Sets of Promising Models

BMA allows us to average over all possible models, containing all possible subsets of features. However, considering many models would require extensive computations, especially when computing the posterior predictive distributions. Such computations would be prohibitively long for a relatively quick screening tool.

Because it was computationally infeasible to consider all possible 2^{100} models, we first chose a subset of the models. For computational efficiency in model selection, this study followed Yeung *et al*. (2005) and used a deterministic search based on an Occam's window approach (Madigan and Raftery, 2004) and the "leaps and bounds" algorithm (Volinsky *et al*., 1997) to identify models with higher posterior probabilities.

2.6 Promoting Feature Sparsity with Iterated BMA

In order to make a more economical screening test, it was important to limit the number of proteins to assay, which required the classification models to use a small subset of features. But Bayesian model averaging was designed to model the model-selection uncertainty and to improve predictive performance (Rafftery, 1995), not to choose a small set of features. To promote feature sparsity, we applied an iterative adaptation of BMA (Yeung *et al*., 2005). This method initially ranks each feature separately by the ratio of between-group to within-group sum of squares (BSS/WSS) (Dudoit *et al*., 2002). For protein j the ratio is

$$\frac{\mathrm{BSS}(j)}{\mathrm{WSS}(j)} = \frac{\sum_i \left[I(Y_i = 0)(\overline{X}_{0j} - \overline{X}_j)^2 + I(Y_i = 1)(\overline{X}_{1j} - \overline{X}_j)^2\right]}{\sum_i \left[I(Y_i = 0)(X_{ij} - \overline{X}_{0j})^2 + I(Y_i = 1)(X_{ij} - \overline{X}_{1j})^2\right]} \quad (6)$$

where $I(\cdot)$ is an indicator function, X_{ij} is the level of protein j under sample i, $\overline{X}_{0j}$ and $\overline{X}_{1j}$ are respectively the average levels of protein j in the normal and cancer groups, and $\overline{X}_j$ is the average level of protein j over all samples.

Ordered by this (BSS/WSS) ranking, iterative BMA runs traditional BMA within each iteration and discards proteins that have low posterior probabilities of relevance, $\Pr(b_j \neq 0|D) < 1\%$, where

$$\Pr(b_j \neq 0|D) = \sum_{M_k \in \Re} \Pr(M_k|D) \quad (7)$$

where $\Re$ is the subset of the considered models $M_1, ..., M_B$ that include protein j. By discarding proteins that have small influence on classification, this iterative procedure keeps only the most relevant proteins.

2.7 Comparing BMA to Other High-Dimensional Classifiers

To compare iterated BMA's classification and generalization performance, we also classified the data using two other dimensionality-reducing methods: a support vector machine (SVM) (Vapnik, 1999) with recursive feature selection (Guyon *et al*., 2002; Zhang *et al*., 2006) and least-angle regression (LAR, a development of LASSO) (Efron *et al*., 2004). All modeling was performed using the R statistical software (version 2.4.1), and specifically the BMA package (version 3.0.3) for iterated BMA, the packages e1071 (version 1.5-16) and R-SVM (http://www.hsph.harvard.edu/bioinfocore/RSVMhome/R-SVM.html) for the SVM with recursive feature selection, and the lars package (version 0.9-5) for least angle regression. We extended the iBMA package to compute the full predictive distributions (as in Figure 5) within cross-validation using an MCMC approach. We sampled from the posterior distributions of the models' regression coefficients and to sample from the posterior predictive distributions, as in Equations 2 and 3.

Table 3. Features chosen by BMA of linear models, normal vs. cancer

Protein	Description
MIF (macrophage migration inhibitory factor)	Inflammation
MMP-9 (matrix metalloproteinase)	Breakdown of extracellular matrix
MPO (myeloperoxidase)	Inflammation, produces HOCl

2.8 Evaluating Classification Performance

The classifiers' performances were analyzed and compared using receiver operating characteristic (ROC) analysis. To estimate generalization performance on future cases, all classifiers were run with leave-one-out cross-validation (LOOCV). Feature selection was performed within each fold of the cross-validation, and feature strength was determined by a feature's selection frequency over the folds.

3 RESULTS

3.1 Normal versus Cancer

Figure 1 shows a plot of the models chosen by Bayesian model averaging of linear models for the classification task of normal vs. cancer. The models are ordered by selection frequency, with the most frequently selected model on the left and the least selected model on the right. Strong, often chosen features will appear as horizontal bands across the plot. Feature coefficients are shown in red for positive coefficient values and blue for negative values. The strongest features are listed in Table 3.

Figure 2 shows the marginal posterior probability distribution functions (PDFs) for the first 9 features. These PDFs of the coefficients were produced by model averaging. The supplementary materials show similar plots for the rest of the features and for the classification tasks of normal tissue vs. benign lesions and normal vs. malignant lesions.

Table 4 shows the classification error, and Figure 3 shows the classifiers' receiver operating characteristic (ROC) curves. All classifiers were run with leave-one-out cross-validation (LOOCV).

The models were also compared in terms of the model size. Figure 4 plots a heatmap of the normalized feature selection frequencies.

Figure 5 plots the full posterior predictive distributions for BMA of logistic models. Similar plots for BMA of linear and probit models are available in the supplementary materials.

3.2 Normal versus Benign, and Benign versus Cancer

Whereas a cancer screening test depends primarily on distinguishing normal tissue from cancer, one can potentially improve the test's specificity by classifying benign lesions. In addition to detecting cancers by classifying normal vs. cancer, we also sought proteins

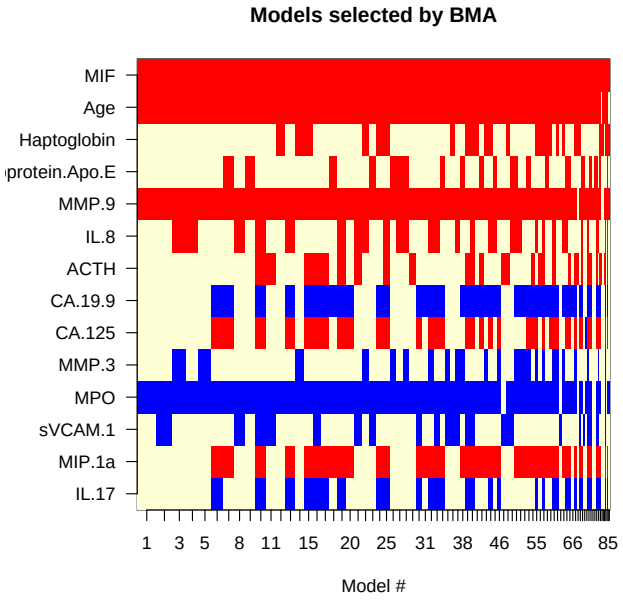


Fig. 1. Models selected by BMA of linear models, normal vs. cancer. Models are ordered by selection frequency, with the best, most frequently selected models on the left and hte weakest, rarest chosen on the right. Coefficients with positive values are shown in red and negative values in blue. Strong, frequently selected features appear as solid horizontal stripes.

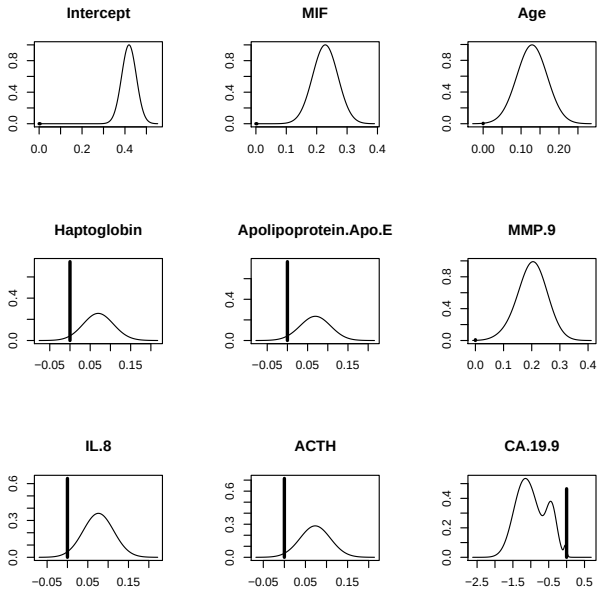


Fig. 2. Marginal posterior distributions of the coefficients, for BMA of linear models, for the classification task of normal vs. cancer. The vertical axis shows the probability values, and the horizontal axis shows the value of the feature j's coefficient β_j. The height of the vertical line segment at $\beta_j = 0$ represents the probability that the coefficient is exactly zero. The nonzero part of the distribution is scaled so that the maximum height is equal to the probability that the coefficient is nonzero.

Table 4. LOOCV classification errors, normal vs. cancer

Model	FN	FP
BMA of linear models	19	8
BMA of logistic models	15	12
BMA of probit models	18	7
SVM with RFS	18	12
LAR	24	5

Classification was performed at the threshold of 0.5 for the classifiers' outputs.

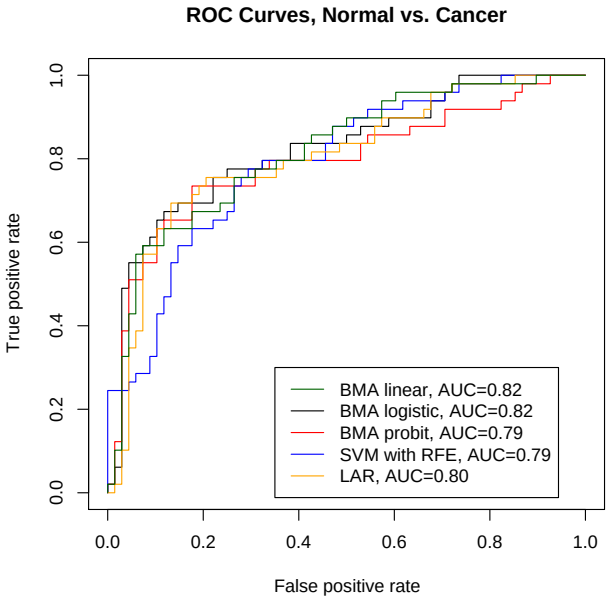


Fig. 3. Normal vs. Cancer, ROC curves. There were no statistically significant differences among the areas under the curves.

that distinguished normal tissue from benign lesions and also benign from malignant lesions.

Figure 6 shows the matrix of selected models for BMA of linear models. The LOOCV classification performance is shown by the ROC curves in Figure 7.

Figure 8 shows the matrix of selected models for BMA of linear models. The classifiers' ROC curves are shown in Figure 9. For both classification tasks, the supplementary materials list the selected proteins and plot the full posterior predictive distributions of the BMA models.

4 DISCUSSION

The group of assayed proteins showed promise in distinguishing normal tissue from lesions. As shown by in Figure 3, for example,

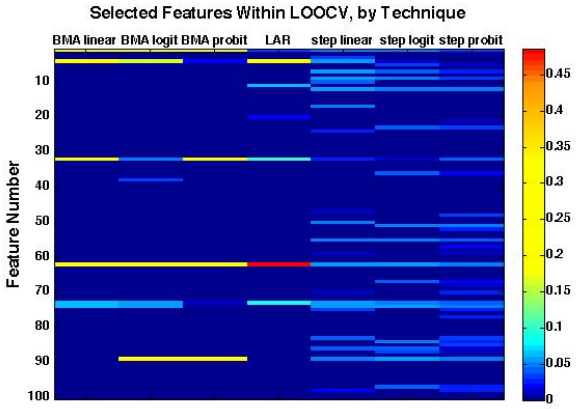


Fig. 4. Heatmap of normalized frequencies of selected features, normal vs. cancer, for the following classifiers: BMA of linear, logistic, and probit models; least angle regression, and linear, logistic, and probit regression with stepwise feature selection. The feature selection frequencies were averaged over all folds of the LOOCV. For comparison across techniques, the frequencies in each column were scaled to sum to one. Less-frequently selected features appear as cooler dark red colors, whereas more frequently selected features appear as hotter, brighter colors. Models that used fewer features appear as dark columns with a few bright bands, whereas models that used more features appear as denser smears of darker bands. Iterated BMA showed a higher feature concentration than did stepwise feature selection.

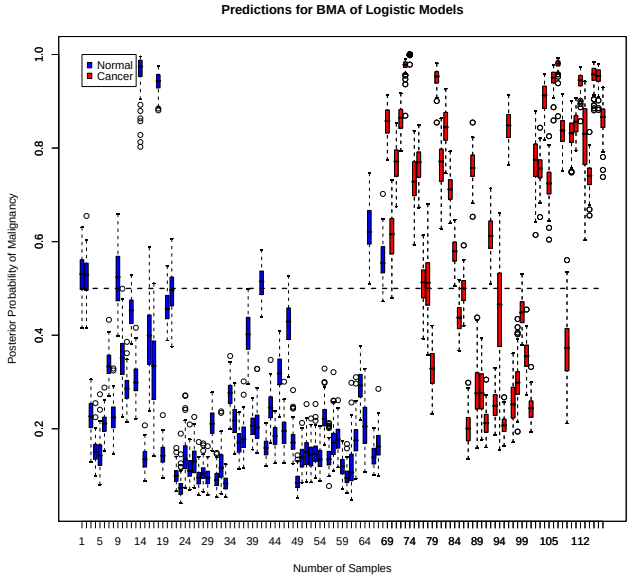


Fig. 5. Posterior probabilities of malignancy, normal vs. cancer. The probabilities came from the iterations of the MCMC chain and are shown as boxplots, with blue for normal subjects and red for subjects with cancer.

BMA of linear models detected 92% of malignancies at a false positive rate of 60% (specificity of 40%). All five classifiers achieved very similar performance, distinguishing normal tissue

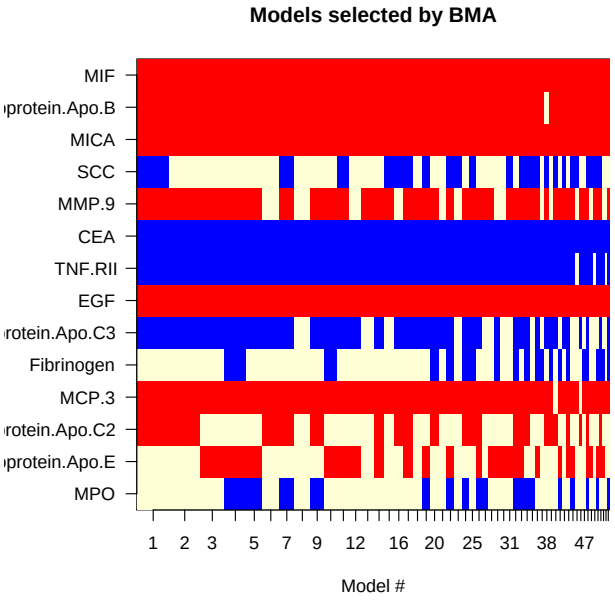


Fig. 6. Models selected by BMA of linear models, normal vs. benign lesions.

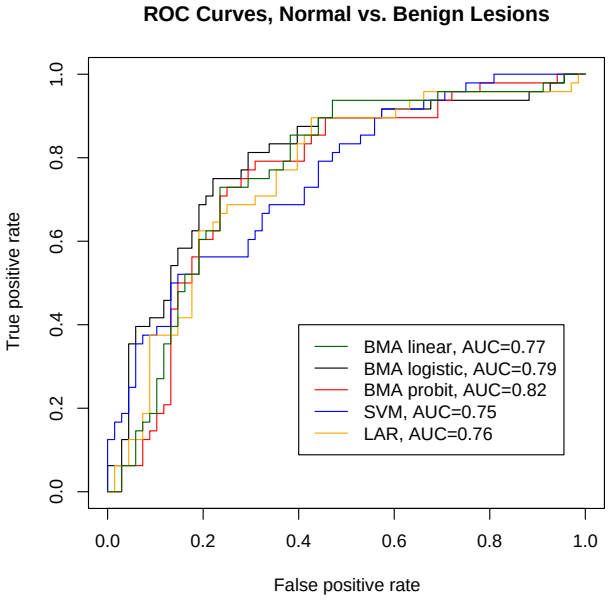


Fig. 7. ROC curves for normal vs. benign. There were no statistically significant differences among the areas under the curves.

from cancer moderately well. The classifiers correctly called approximately 144 of the 171 cases (Table 4) and had an area under the ROC curve of approximately $AUC = 0.80$ (Figure 3).

Although the selected serum proteins (Table 3 and supplementary materials) were moderately successful at detecting the presence

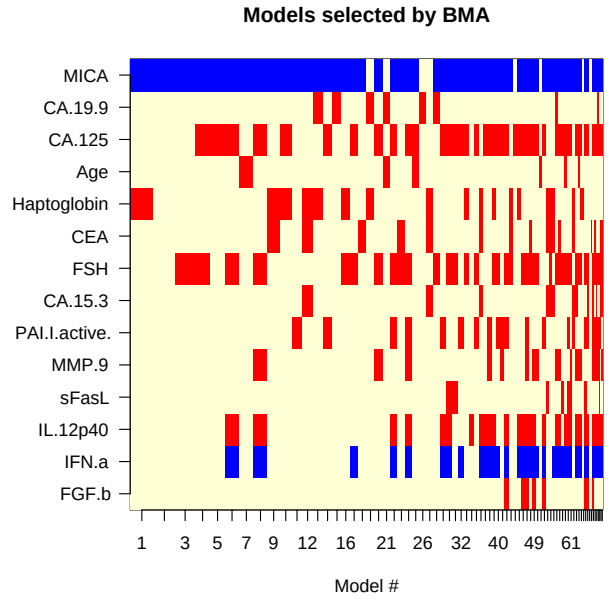


Fig. 8. Models selected by BMA of linear models, ordered by frequency of selection, for the classification task of benign vs. cancer. This part of the data set had only one consistently selected feature: MICA.

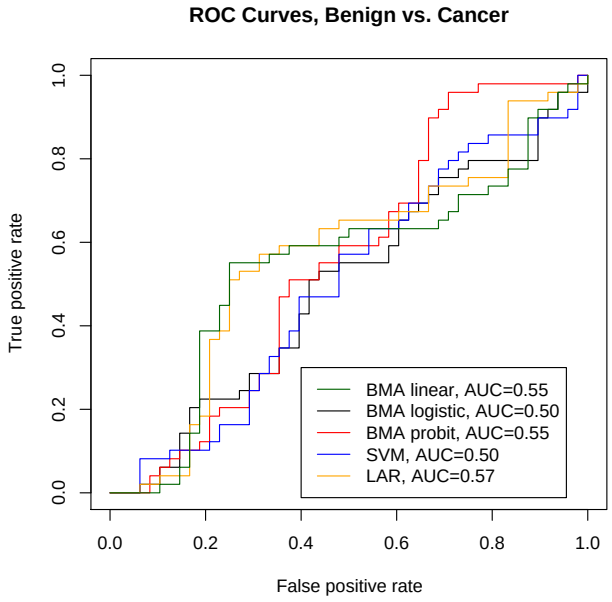


Fig. 9. ROC curves for benign vs. cancer. None of the classifiers were able to distinguish benign from malignant lesions.

of cancer, their classification performance within benign lesions suggested that the proteins were probably more indicative of secondary effects rather than specific for cancer. Proteins for both benign and normal lesions resulted in very similar classification performances, with $AUC = 0.80$ for malignant lesion and $AUC = 0.78$ for benign lesions (Figures 3 and 7). Nearly identical classification results suggest that the proteins may signal more vague states of biological or immunological stress. A good candidate for the dominating biological effect is inflammation, since the top protein selected for both normal vs. cancer and normal vs. benign was macrophage migration inhibitory factor (MIF), which has been shown to be active in inflammation. The protein composition was very similar for benign and malignant lesions, as seen by the classifiers' inability to distinguish benign from malignant lesions ($AUC = 0.52$). Although this classification task did have one consistently chosen feature, human major histocompatibility complex class I chain-related A (MICA), it did not allow for good predictive performance (Figure 9).

In fact this question of secondary effects raises the general concern about identifying circulating biomarkers: are the observed marker candidates actually relevant to the disease under consideration? This concern pertains especially to general and common secondary effects, such as immune response. Although it is difficult to address this concern with certainty, helpful study designs would control for known likely secondary causes and collect enough samples to average over unintended secondary causes. Longitudinal studies would also lessen the effect of transient secondary causes.

To quantify and compare classification performances, we used ROC analysis, which fairly compares classifiers that may be operating at different sensitivities due to arbitrary decision thresholds applied to the classifiers' output values. Although our data set comprised three classes (normal, benign, and cancer), current ROC methods required us to split an inherently three-class classification problem into three different two-class tasks: normal vs. benign, normal vs. cancer, and benign vs. cancer. The field of ROC analysis is still in development for the three-class problem; no consensus has yet been reached about how to quantitatively score the resulting six-dimensional ROC hypersurface. However, for other methods of classifier comparison, such as the generalized Brier score or discrete counts of classification errors, full three-class models could have been used.

This study's classification results came from a group of 98 serum proteins (Table 1), which is relatively small sample of all detectable serum proteins. Future studies may identify other proteins with stronger relationships to breast cancer, or predictive performance could also benefit from a much larger group of weakly associated proteins. Perhaps it will become more feasible to screen large populations with protein-based tests that require a larger set of proteins when the design and manufacturing costs lower for microfluidics chips. Such arrays would simplify the process of automating blood tests in a high-throughput fashion. However, with current assay technology and cost-benefit analysis of screening programs, the fixed cost per protein assayed essentially limits the number of proteins used for screening. To lower screening costs, we chose small subsets of the features via feature-selection methods. As seen in Figure 4, BMA and least-angle regression were able to classify well using a far smaller set of features than those chosen by stepwise feature selection.

By creating redundant features, feature correlations impede many feature-selection techniques. For stochastic feature-selection methods that select for informative, non-redundant features, two highly correlated features are each likely to be chosen in alternation. Similarly, a cluster of highly correlated features causes the feature

selection technique to spread the feature selection rate among each feature in the cluster, essentially diluting each feature's selection rate. Severe dilution of selection rates can cause none of the cluster's features to be chosen. Future work will entail adding cluster-based methods to the iterated BMA algorithm.

The true benefit of protein-based breast cancer screening will depend on its relationship to existing imaging-based screening. The proteins will boost diagnostic performance only if they provide complementary and non-redundant information with the clinical practice of mammograms, sonograms, and physical examination. The relationship of imaging and protein screening remains to be determined in future work.

5 CONCLUSION

We have performed feature-selection and classification techniques to identify blood serum proteins that are indicative of breast cancer. The best features to detect breast cancer were MIF, MMP-9, and MPO. While the proteins could distinguish normal tissue from cancer and normal tissue from benign lesions, they could not distinguish benign from malignant lesions. Since the same protein (MIF) was chosen for both normal vs. cancer and normal vs. benign lesions, it is likely that this protein plays a role in the inflammatory response to a lesion, whether benign or malignant, rather than in a role specific for cancer. While the current set of proteins show promise in detecting breast cancer, their true usefulness in a screening program remains to be seen in their integration with current imaging-based screening practices.

ACKNOWLEDGEMENT

The work was supported in part by the NIH (NIH CA 84955 and R01 CA-112437-01) and the U.S. Army Breast Cancer Research Program (Grant No. W81XWH-05-1-0292).

REFERENCES

American Cancer Society (2007) Cancer statistics for 2007. Website: http://www.cancer.org/docroot/stt/stt_0.asp.

Berger,J.O. and Pericchi,L.R. (2001) Objective Bayesian methods for model selection: Introduction and comparison (with discussion). In *Model Selection* (Lahiri,P. ed.), IMS, Beechwood, OH, pp. 135-207.

Cady,B., and Michaelson,J.S. (2001) The life-sparing potential of mammographic screening, *Cancer*, **91**, 1699-1703.

Chipman,H.A. *et al*. (2001) The practical implementation of Bayesian model selection (with discussion). In *Model Selection* (Lahiri,P. ed.), IMS, Beechwood, OH, pp. 65-134.

Clyde,M. and George,E.I. (1994) Model uncertainty. *Stat. Sci.*, **19**, 81-94.

Draper,D. (1995) Assessment and propagation of model uncertainty (with discussion). *J. Roy. Stat. Soc. Ser. B*, **57**, 903-905.

Dubsky,P.C. *et al*. (2002) Young age as an independent adverse prognostic factor in premenopausal patients with breast cancer, *Clin. Breast Cancer*, **3**, 65-72.

Dudoit *et al*. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77-87.

Efron,B., Hastie,T., Johnstone,I., and Tibshirani,R. (2004) Least angle regression. *Ann. Statist.*, **32**, 407-499.

Gorelik,E. *et al*. (2005) Multiplexed immunobead-based cytokine profiling for early detection of ovarian cancer, *Cancer Epidemiol. Biomarkers Prev.*, **14**, 981-987.

Gruvberger,S. *et al*. (2001) Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.*, **61**, 5979-5984.

Guyon,I. *et al*. (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, **46** 389-422.

Hodges,J.S. (1987) Uncertainty, policy analysis and statistics (with discussion). *Stat. Sci.*, **2**, 259-275

Hoeting,J.A. *et al*. (1999) Bayesian model averaging: a tutorial. *Stat. Sci.*, **14**, 382-37.

Hye,A. *et al*. (2006) Proteome-based plasma biomarkers for Alzheimer's disease, *Brain*, **129**, 3042-3050.

Jemal,A. *et al*. (2007) Cancer statistics, *CA Cancer J. Clin.*, **57**, 43-66.

Kerlikowske,K. *et al*. (1995) Efficacy of screening mammography. A meta-analysis, *JAMA*, **273**, 149-154.

Kreunin,P., Yoo,C., Urquidi,V., Lubman,D.M., Goodison,S. (2007) Proteomic profiling identifies breast tumor metastasis-associated factors in an isogenic model. *Proteomics*, **7**, 299-312.

Ferrini,R. *et al*. (1996) Screening mammography for breast cancer: American College of Preventive Medicine practice policy statement, *Am. J. Prev. Med.*, 12, 340-341.

Lacey,J.V., Jr. *et al*. (2002) Recent trends in breast cancer incidence and mortality, *Environ. Mol. Mutagen.*, **39**, 82-88.

Lee,K.E. *et al*. (2003) Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19**, 90-97.

Madigan,D.M. and Raftery,A.E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statis. Assoc.*, **89**, 1535-1546.

Malkas,L.H. *et al*. (2006) A cancer-associated PCNA expressed in breast cancer has implications as a potential biomarker. *PNAS*, **103**, 1947219477.

Martin,K.J. *et al*. (2000) Linking gene expression patterns to therapeutic groups in breast cancer. *Cancer Res.*, **60**, 2232-2238.

Meyer,J.E. *et al*. (1984) Occult breast abnormalities: percutaneous preoperative needle localization, *Radiology*, **150**, 335-337.

Perou,C.M. *et al*. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, *Proc. Natl Acad. Sci. USA*, **96**, 9212-9217.

Polascik,T.J. *et al*. (1999) Prostate specific antigen: a decade of discovery what we have learned and where we are going, *J. Urol.*, **162**, 293-306.

Rafftery,A.E. (1995) Bayesian model selection in social research (with Discussion). In Marsden,P.V. (ed.) *Sociological Methodology 1995*, Blackwell, Cambridge, MA, pp. 111-196.

Rafftery,A.E. *et al*. (1995) Accounting for model uncertainty ni survival analysis improves predictive performance (with Discussion). In Bernardo,J.M., Berger,J.O., Dawid,A.P., and Smith,F.M. (eds.), *Bayesian Statistics 5*, Oxford University Press, Oxford, UK, pp. 323-349.

de la Rochfordiere,A., and Asselain,B. (1993) Age as prognostic factor in premenopausal breast carcinoma, *Lancet*, **341**, 1039-1043.

Rosenberg,A.L. *et al*. (1987) Clinically occult breast lesions: localization and significance, *Radiology*, **162**, 167-170.

Sha,N. *et al*. (2004) Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, **60**, 812-819.

Simpson,H.W. *et al*. (1988) Genesis of breast cancer is in the premenopause, *Lancet*, **332**, 74-76.

Sorlie,T. *et al*. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **98**, 10869-10874.

Tabar,L. *et al*. (1995) Efficacy of breast cancer screening by age. New results from the Swedish Two-County Trial, *Cancer*, **75**, 2507-2517.

van't Veer,L.J. *et al*. (2001) Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, **415**, 530-536.

Vapnik,V. (1999) The Nature of Statistical Learning Theory. Springer-Verlag.

Volinsky,C. *et al*. (1997) Bayesian model aaveraging in proportional hazard models: assessing stroke risk. *Appl. Statis.*, **46**, 433-488.

Wang,T.J. *et al*. (2006) *N. Engl. J. Med.*, **355**, 2631-2369.

West,M. *et al*. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462-11467.

Yankaskas,B.C. *et al*. (1988) Needle localization biopsy of occult lesions of the breast, *Radiology*, **23**, 729-733.

Yeung,K.Y. *et al*. (2005) Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bionformatics*, **21**, 2394-2402.

Zajchowski, D.A. *et al*. (2001) Identification of gene expression profiles that predict the aggressive behavior of breast cancer cells. *Cancer Res.*, **61**, 5168-5178.

Zhang,X. *et al*. (2006) Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*, **7**, 197.